Kristin Tøndel

Rational design of protein inhibitors using molecular modelling and multivariate analysis

Doctoral thesis
for the degree of doktor ingeniør

Trondheim, March 2004

Norwegian University of Science and Technology
Department of Chemistry

NTNU

ISBN 82-471-6329-2
ISBN 82-471-6328-4(electronic ver.)

Preface

My work includes all steps in a rational drug design process, from protein structure prediction with homology modelling, to the actual construction and ranking of drug candidate structures from proposed active functional groups. In this thesis, I have tried to give a brief overview of the methods that are most relevant to my work, as well as a summary of the most important results presented in the papers. The topics of the thesis are:

1. Protein structure modelling and prediction of homology model accuracy.
2. Mapping of the properties of protein binding sites and identification of interaction sites for selective inhibitors.
3. *De novo* ligand design.
4. Molecular docking and estimation of binding energies between proteins and small-molecular ligands.

First, a brief introduction to rational drug design is given, followed by an introduction to the computational methods used in this work. For some of the methods, books or reviews are cited instead of the original publications. Some background information about the molecular systems studied in this work is also given. Following this introduction to the background theory, I give an overview of my work and show how the different parts of my work are linked together. Then a short presentation of each part of the work is given. In order to give the reader a more coherent reading of the text, I have chosen to write the thesis as a complete manuscript, independent of the papers. This makes it possible for the reader to get an overview of the work and the main results without having to read all the details in the papers. The papers are given at the end of the thesis, and give a more detailed description of each part of the work. In this way, I aim to increase the number of interested readers, because those who are not interested in the details, but just an overview, can skip reading the papers. Some parts of the text are therefore given both in the papers and in the thesis. In the text, the papers included in the thesis are referred to as Paper I-VII. These are not included in the list of references. The results presented in Figure 4 and 8 in Paper I are shown in more detail in Appendix 3 and 4. All results except the results presented in Chapter 5.3.3 and Appendix 1 and 2 and the docking results presented in Table 5.1 (page 57) are published in the papers. All scripts written in the present work can be obtained upon request. Some of the scripts are written in Scientific Vector Language (SVL). This is the scripting language contained in the commercial software package Molecular Operating Environment™ (MOE), provided by the Chemical Computing Group, Inc. Hence, usage of these scripts requires access to this software package.

Acknowledgements

A special thanks to Dr. Finn Drabløs at the Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology and Endre Anderssen at the Department of Chemistry, Norwegian University of Science and Technology for their important contributions to large parts of the work presented here. Dr. Astrid Hilde Myrset at Amersham Health is thanked for providing useful tips on how to structure a thesis. Thanks also to my supervisor during the last two years of my Ph.D. period, Prof. Per-Olof Åstrand for organising and encouragement.

Thanks to Prof. J. Andrew McCammon and his research group at the Department of Chemistry and Biochemistry at the University of California, San Diego (UCSD), for allowing me to visit their group for seven months. Dr. Chung F. Wong and Dr. Jens Erik Nielsen have made important contributions to my work. During my stay, I studied the interactions between the receptor kinase fibroblast growth factor receptor 1 (FGFR1) and a known inhibitor, and worked on identification of more effective and selective inhibitors of this protein kinase. I also participated in the development of a pipeline for automatic all-against-all homology modelling from a multiple sequence alignment. In the same project, an automatic method for homology model quality verification by calculation of root mean square deviations (RMSDs) and differences in inter-residue contact areas between the models and experimental structures of the targets was developed. The produced dataset was used in the development of a method for prediction of homology model accuracy.

I also thank Prof. Anders Sundan and Dr. Magne Børset at the Department of Cancer Research and Molecular Biology, Norwegian University of Science and Technology for helpful discussions, and Dr. Randi Nygaard at the Children's Clinic at Trondheim University Hospital for helping me to gain insight into the field of cancer research.

Elena L. Vodovozova and co-workers at the Laboratory of Lipid Chemistry and Biochemistry at the Institute of Bioorganic Chemistry, Russian Academy of Sciences are thanked for providing us with a dataset containing experimental binding affinities of 34 carbohydrate ligands to cancer cells.

The research group at the Department of Chemistry, section for Physical Chemistry at the Norwegian University of Science and Technology, Terje Bruvoll in particular, is thanked for creating an encouraging and inspiring work environment.

Finally, I thank the Norwegian Research Council for financial support.

Summary

Cancer cells utilise signalling cascades involving protein kinases for their growth and survival. Hence, design of protein inhibitors that block the function of these signalling proteins is interesting for the development of new cancer therapies. This can be achieved through rational drug design. The purpose of this work was to develop drug design methods that can aid the discovery of selective drugs and to utilise these methods to design drugs that block the function of proteins involved in cancer development. This work has focused on development of drug design methods that can be used with protein structure models made by homology modelling, since this will significantly increase the number of protein targets for which the methods can be used. In this context, a review about the use of homology-based modelling in rational drug design was published. The design part has focused on design of selective inhibitors of protein kinases, in particular Tyrosine kinase 2 (Tyk2), a member of the Janus kinase (Jak) family of protein kinases. The interactions between the receptor kinase fibroblast growth factor receptor 1 (FGFR1) and a known inhibitor have also been studied, and several improvements of the inhibitor have been suggested, based on results from computational sensitivity analysis. Both in the case of Tyk2 and FGFR1, focus has been on inhibiting the binding of adenosine triphosphate (ATP) to the tyrosine kinase domain of the proteins. The interactions between E-selectin and a set of carbohydrates and peptide ligands were also studied with computational docking. The results from this study provide insight into some of the limitations of docking methods.

In order to analyse the relationship between the target-template similarity and the accuracy of the obtained homology model, a large number of homology models for protein kinase structures were generated, and the accuracy of the homology models was evaluated by comparison to available experimental structures of the targets. Based on the obtained data, a new method for prediction of homology model accuracy with multivariate regression was developed, that predicts the model accuracy directly from the amino acid sequence alignment. This method can be used to assure that the optimal templates are chosen, and for identification of regions of the protein structure that are difficult to model, as well as errors in the alignment of the proteins. Here, this method has been applied to the protein kinase family, but the same approach can be used for other protein families.

A new method for analysis of protein binding site properties, called Protein Alpha Shape Similarity Analysis (PASSA), and a new gaussian-based docking method suitable for use with homology modelled protein structures have been developed. Both methods use gaussian functions to represent atomic properties. This smooth representation makes them relatively robust against small structural errors. PASSA has been shown to be a useful method for identification of regions in a protein binding site that can be utilised to achieve selective binding of ligands to the protein. Interaction sites identified by PASSA to be important for selectivity have been shown to correspond to functional groups of known, selective inhibitors. The gaussian-based docking method developed here is relatively fast, and well suited for virtual screening, where the purpose is to seek out a set of promising drug candidates from a large amount of ligand structures. However, the accuracy of our docking method cannot be compared to that of other methods that use fewer approximations. In contrast to many other docking methods, our docking method predicts hydrophobic interactions better than hydrophilic interactions.

PASSA was used to suggest functional groups for a selective inhibitor of Tyk2. The results from this study were used further in a screening of the database of the National Cancer Institute (NCI) for possible Tyk2 inhibitors. The proposed functional groups were also combined into drug candidates by *de novo* ligand design. The gaussian-based docking method developed here was applied to rank the drug candidate molecules resulting from the database screening and *de novo* ligand design according to binding to Tyk2. The selectivity of the compounds was tested by computational docking in seven other protein kinase structures. The results from the docking of the compounds from the NCI database were compared to the results obtained using another docking

method, MOE-Dock. The two docking methods ranked the structures differently, but produced the same conclusion, namely that none of the compounds in the NCI database can inhibit Tyk2 selectively. One compound was found to inhibit Tyk2 and insulin receptor tyrosine kinase selectively, and five of the drug candidates from the *de novo* ligand design seem promising as selective Tyk2 inhibitors. These results have to be verified experimentally, of course.

PASSA has also been used to model selectivity within the protein kinase family. In this way, the PASSA method may be used quantitatively to predict activities for a number of ligands within a set of closely related protein targets. This makes PASSA a promising method in screening for side effects. This method also allows for effective visualisation of the molecular basis for selectivity.

The results presented here indicate that methods utilising gaussian functions to describe molecular properties have many applications in structure-based drug design, and will be useful supplements to other methods. These methods seem especially useful in the initial stages of a drug design process, when computational efficiency and robustness are most important.

List of publications

Paper I. Kristin Tøndel, Prediction of homology model quality with multivariate regression, *Journal of Chemical Information and Computer Sciences,* Submitted.

Paper II. Kristin Tøndel, Endre Anderssen and Finn Drabløs, Protein Alpha Shape Similarity Analysis (PASSA): A new method for mapping protein binding sites. Application in the design of a selective inhibitor of Tyrosine kinase 2, *Journal of Computer-Aided Molecular Design*, 2002, 16, 831-840.

Paper III. Heather Wieman, Kristin Tøndel, Endre Anderssen and Finn Drabløs, Homology-based modelling of targets for rational drug design, *Mini-Reviews in Medicinal Chemistry*, 2004, In Press.

Paper IV. Kristin Tøndel, Chung F. Wong and J. Andrew McCammon, Computational analysis of the interactions between the angiogenesis inhibitor PD173074 and fibroblast growth factor receptor 1, *Journal of Theoretical and Computational Chemistry*, 2003, 2, 43-56.

Paper V. Kristin Tøndel, Endre Anderssen and Finn Drabløs, A new gaussian-based docking method suitable for use with homology modelled proteins, *Journal of Computer-Aided Molecular Design*, Submitted.

Paper VI. Kristin Tøndel and Finn Drabløs, Design of selective inhibitors of Tyrosine kinase 2, *Journal of Medicinal Chemistry*, Submitted.

Paper VII. Endre Anderssen and Kristin Tøndel, Application of Protein Alpha Shape Similarity Analysis (PASSA) in modelling selectivity, *Journal of Computer-Aided Molecular Design,* Submitted.

List of abbreviations

2D	Two-dimensional
3D	Three-dimensional
Abl kinase	Abelson kinase
ADMET	Absorption, distribution, metabolism, excretion and toxicity
APROPOS	Automatic PROtein Pocket Search
ATP	Adenosine triphosphate
CAD	Contact Area Difference
CNTF	Ciliary neurotrophic factor
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity index analysis
DNA	Deoxyribonucleic acid
DPLSR	Discriminant Partial Least Squares Regression
EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
Epo	Erythropoietin
FGF	Fibroblast growth factor
FGFR	Fibroblast growth factor receptor
GB/SA	Generalised Born Surface Area method
G-CSF	Granulocyte-specific colony-stimulating factor
GH	Growth hormone
GM-CSF	Granulocyte-macrophage colony-stimulating factor
Hck	Haematopoetic cell kinase
IC_{50}	Inhibitory concentration (50%)
ICM	Internal coordinate mechanics
IFN	Interferon
IFREDA	Internal coordinate mechanics (ICM)-flexible receptor docking algorithm
IGF	Insulin-like growth factor
IL	Interleukin
Jak	Janus kinase
JH	Jak homology
Lck	Lymphocyte-specific kinase
LIF	Leukaemia inhibitory factor
MC	Monte Carlo simulation
MCSS	Multiple copy simultaneous search
MD	Molecular dynamics
MM	Molecular mechanics
MOE	Molecular Operating Environment
NCI	National Cancer Institute
NMR	Nuclear magnetic resonance
OLS	Ordinary Least Squares
OWFEG	One window free energy grid
PAM	Point Accepted Mutation
PASS	Putative Active Sites with Spheres
PASSA	Protein Alpha Shape Similarity Analysis
PB/SA	Poisson-Boltzmann Surface Area method
PC	Principal Component
PCA	Principal Component Analysis
PDB	RCSB Protein Data Bank
PDECGF	Platelet-derived endothelial cell growth factor

pdf	Probability density function
PDGF	Platelet-derived growth factor
PDGFR	Platelet-derived growth factor receptor
PKA	Protein kinase A
PKC	Protein kinase C
PLP	Pairwise linear potentials
PLS	Partial Least Squares
PMF	Potential of Mean Force
QSAR	Quantitative Structure-Activity Relationship
RMSD	Root mean square deviation
RTK	Receptor tyrosine kinase
SegMod	Segment Match Modelling
SH	Src homology
$SiaLe^a$	Sialyl Lewis a
$SiaLe^x$	Sialyl Lewis x
SLIDE	Screening for Ligands by Induced-fit Docking, Efficiently
SPECITOPE	The "Specific Epitope" docking program
STAT	Signal Transducers and Activators of Transcription
TGF	Transforming growth factor
TNF	Tumour necrosis factor
Tyk2	Tyrosine kinase 2
UHBD	The University of Houston Brownian Dynamics program
vdW	van der Waals
VEGF	Vascular endothelial growth factor
VEGFR	Vascular endothelial growth factor receptor

Amino acid names:

ALA	A	Alanine
ARG	R	Arginine
ASN	N	Aspargine
ASP	D	Aspartic acid
CYS	C	Cysteine
GLU	E	Glutamic acid
GLN	Q	Glutamine
GLY	G	Glycine
HIS	H	Histidine
ILE	I	Isoleucine
LEU	L	Leucine
LYS	K	Lysine
MET	M	Methionine
PHE	F	Phenylalanine
PRO	P	Proline
SER	S	Serine
THR	T	Threonine
TRP	W	Tryptophan
TYR	Y	Tyrosine
VAL	V	Valine

Contents

Preface....1
Acknowledgements....3
Summary....4
List of publications....6
List of abbreviations....7
Contents....9
1 Introduction....11
2 Computational methods....14
2.1 Basic principles of molecular modelling....14
2.1.1 Molecular mechanics....14
2.1.2 Conformational searching....15
2.1.3 Protein structure prediction....16
2.1.4 Binding free energy estimation....17
2.1.5 Surface area calculations....19
2.2 Multivariate regression....20
2.3 Rational drug design methods....21
2.3.1 Homology modelling methods....21
2.3.2 Methods for verification of the accuracy of protein structure models....22
2.3.3 Methods for mapping protein binding site properties....24
2.3.4 Computational docking methods....25
2.3.5 Score functions for computational docking....28
2.3.6 *De novo* ligand design....30
3 Molecular systems....32
3.1 Protein kinases....32
3.1.1 Janus kinases....33
3.1.2 Fibroblast growth factor receptor....34
3.2 Lectins....34
4 Summary of the papers....36
5 Case studies....39
5.1 Verification and prediction of homology model accuracy....39
5.2 Mapping protein binding site properties....41
5.3 Modelling interactions between proteins and drug candidates....46
5.3.1 Computational analysis of the interactions between the angiogenesis inhibitor PD173074 and fibroblast growth factor receptor 1....46
5.3.2 A new gaussian-based docking method suitable for use with homology modelled proteins....48
5.3.3 Computational docking of carbohydrate ligands and peptides in E-selectin....52
5.4 Design of selective inhibitors of Tyrosine kinase 2....56
5.4.1 Method testing and verification of the structural model....56
5.4.2 Database screening and *de novo* ligand design....57
5.5 Application of Protein Alpha Shape Similarity Analysis (PASSA) in modelling selectivity....61
6 Discussion....64
7 Conclusions....66
8 Future perspectives....67

Appendix 1. Computational details of the docking of carbohydrate ligands and peptides in E-selectin.....68
Appendix 2. Results from the docking of carbohydrate and peptide ligands in E-selectin.....69
Appendix 3. Multiple sequence alignment and alignment score profiles for the kinases studied in Paper I.....71
Appendix 4. Multiple sequence alignment and regression coefficients from Paper I.....74
References.....77
Paper I-VII

1 Introduction

Rational drug design has been shown to have a large potential as a tool for the pharmaceutical industry, saving both time and money, compared to the conventional "trial and error" approach (Sawyer, 2001). The purpose is to design a drug (normally a small molecule) that moderates the normal function of a target (usually a protein) in a selective and normally reversible way, using computational methods. In addition to this, several physical criteria have to be met, related to production, uptake, degradation, etc. The large amount of data from large-scale genome oriented projects, such as the human genome project (McPherson *et al.*, 2001), eases both the identification of suitable targets and the actual drug design process. Information about genome sequences, regulatory networks and metabolic pathways, combined with biological and medical data creates the basis for identifying optimal drug targets (Burley *et al.,* 1999). Access to high-quality three-dimensional (3D) structures of these targets is a good starting point for rational drug design. In addition to the traditional experimental methods for generating models of protein 3D structures (X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy), fast and efficient computational methods, such as homology modelling, have been developed. This increases the number of possible targets that can be analysed by rational drug design significantly. However, since most existing drug design methods are trained on and developed for use with experimental 3D structures of proteins, there is a need for more robust methods that account for the additional error resulting from using homology models of the protein structures. Methods that include protein flexibility in the calculations are more robust against small errors in the protein structure models than methods that treat the receptor as a rigid structure. The fact that ligand binding can induce conformational changes in the protein structure also makes the development of efficient methods to account for protein flexibility important.

In general, computational drug design can be divided into two main approaches. One class of methods, the Quantitative Structure-Activity Relationship (QSAR) methods, starts from a set of known drug molecules, and tries to predict the activities of new compounds based on the relationship between the properties and the activities of these known drugs (Nikolova and Jaworska, 2003). Another main approach starts from a structural model of the system under consideration (typically a target protein) and tries to design new drugs based on analysis of this protein structure and its interactions with various drug candidates. The work presented here has focused on the second category of drug design methods, and in this text the terms "rational drug design" and "virtual drug design" refer to this approach.

QSAR methods use variables representing the properties of known drug molecules (two-dimensional or three-dimensional), and correlate these to known activities of these molecules using data analysis tools like multivariate regression. The obtained regression models can be used to predict the activities of new drug candidates. Comparative molecular field analysis (CoMFA) (Cramer *et al.*, 1988) and comparative molecular similarity index analysis (CoMSIA) (Klebe *et al.*, 1994) are examples of widely used 3D-QSAR methods.

The virtual drug design process can be divided into four main steps:

1. Analysis of protein binding site properties and identification of possible interaction sites for drug candidates.
2. Suggestion of ligand functional groups that can utilise the identified interaction sites.
3. Database screening for existing drugs having the desired properties and *de novo* design of ligands containing the suggested functional groups.
4. Ranking of drug candidates by computational docking and binding affinity prediction.

In order to design a ligand for a given target, the first step is to analyse the properties of the protein binding site and identify possible interaction sites for ligands. A variety of methods exist for

mapping the properties of protein binding sites. Most of these methods utilise calculations of interaction energies between the protein and small, molecular probes (Sotriffer and Klebe, 2002). Once possible interaction sites for a selective inhibitor have been identified, databases of already existing drugs can be searched in order to find a drug molecule that fits the receptor binding site (Miller, 2002). A number of such databases exist, such as the Cambridge Structural Database (Allen *et al.*, 1983), the database of the National Cancer Institute (NCI) (http://cactus.nci.nih.gov/), the Available Chemicals Directory (MDL Information Systems) and PDBsum (http://www.biochem.ucl.ac.uk/bsm/pdbsum/), which includes a database of ligands from the RCSB Protein Database (PDB) (Berman *et al.*, 2000). Drug candidate molecules can also be generated from proposed ligand functional groups by a process called *de novo* ligand design. Several different approaches to linking the functional groups together, or "growing" ligand structures from one or more "seed" fragments have been developed (Schneider and Böhm, 2002; Anderson, 2003). The ligand structures are typically fitted into the protein binding site using an energy function. Promising drug candidates resulting from database searching or *de novo* ligand design are then ranked according to success of binding to the target protein by estimating binding affinities. This can be done with computational docking (Bajorath, 2002; Halperin *et al.*, 2002; Lyne, 2002; Taylor *et al.*, 2002; Brooijmans and Kuntz, 2003).

Considering only the target protein may be a mistake. Side effects have led to the withdrawal of many drugs from late stage testing (Smith, 2002). Hence, to achieve selectivity and avoid side effects, analysis of related binding sites and estimation of binding affinities between the drug candidates and proteins related to the target is also important. Homology modelling (modelling of a protein structure based on an experimental structure of a related protein) (Bajorath *et al.*, 1993; Sánchez and Sali, 1997; Marti-Renom *et al.*, 2000) makes this possible, since a large number of protein structure models can be obtained. Databases of protein structures can also be searched for proteins that have structural similarities to the target, even though they are not evolutionary related (Holm and Sander, 1993; Murzin *et al.*, 1995; Holm and Sander, 1997; Holm and Sander, 1998 a, b; LoConte *et al.*, 2000; Pearl *et al.*, 2000). In this way, possible side effects due to structural similarities to the target protein can be detected. Homology modelling is also useful in the development of personalised drugs, that is, drugs especially suited for individuals with a specific mutation in the genes coding for a target protein. The fact that several different variants of a target protein are possible implies that the same drugs may not be optimal for all individuals. Since experimental determination of the structures of all variants of a protein is impractical, homology modelling becomes an important tool.

Because of the many approximations to the real biological system used in rational drug design, the results from a rational drug design process have to be verified experimentally before reliable conclusions can be drawn about the activities of these compounds. In particular, the solvent effect and the effects of target and ligand flexibility are often imprecisely described. High accuracy is usually associated with a high computational cost. The computational time required for reliable results increases dramatically with increasing size and number of rotatable bonds of the ligand. The receptor is usually treated as a rigid structure in rational drug design methods. Some approaches include protein flexibility, but these methods are usually very time consuming. Protein flexibility in drug design has recently been reviewed (Carlson and McCammon, 2000; Carlson, 2002; Teodoro and Kavraki, 2003; Wong and McCammon, 2003).

Computational docking involving large and flexible compounds, such as peptides and complex carbohydrate ligands, represents a great challenge. In addition, empirical methods are sensitive to deviations between the target system and the structures used to train the methods. For example projects involving membrane proteins suffer from the low number of available X-ray structures caused by the difficulties in crystallising the proteins (Dahl *et al.*, 2002). However, in spite of these limitations, rational drug design is showing an increasing importance in pharmaceutical research, and much research effort is devoted to the development of new and more effective methods. Several

reviews on rational drug design are available (Apostolakis and Caflisch, 1999; Finn and Kavraki, 1999; Klebe, 2000; Bajorath, 2001; Sawyer, 2001; Stahura and Bajorath, 2002; Anderson, 2003).

2 Computational methods

Rational drug design includes both molecular modelling and multivariate data analysis. Some background theory on these disciplines will therefore be given, prior to a more detailed description of the methods used in rational drug design.

2.1 Basic principles of molecular modelling

2.1.1 Molecular mechanics

Molecular mechanics (MM) is a method for calculation of the structure and energy of molecules based on force field models. In a very simplified sense, MM treats a molecule as a collection of weights connected by springs, where the weights represent the nuclei and the springs represent the bonds.

A force field consists of a collection of atom types that define the atoms in a molecule, parameters for bond lengths, bond angles, etc. and equations for calculation of the energy of a molecule. In a force field, a given element may have several atom types, depending on what kind of functional group it is a part of.

The total energy of a molecule is a sum of several energy terms that are calculated independently. Examples of energy terms include energies associated with bond stretching, bond bending, torsional strain, van der Waals interactions (vdW) and electrostatic interactions (ele). These equations define the potential energy, E_{pot}, of a molecule:

$$E_{pot} = E_{stretch} + E_{bend} + E_{torsion} + E_{vdW} + E_{ele} \quad (2.1)$$

The different terms of Equation 2.1 are illustrated in Figure 2.1. Sometimes additional terms, such as stretch-bend coupling terms, are added to Equation 2.1 (Leach, 2001).

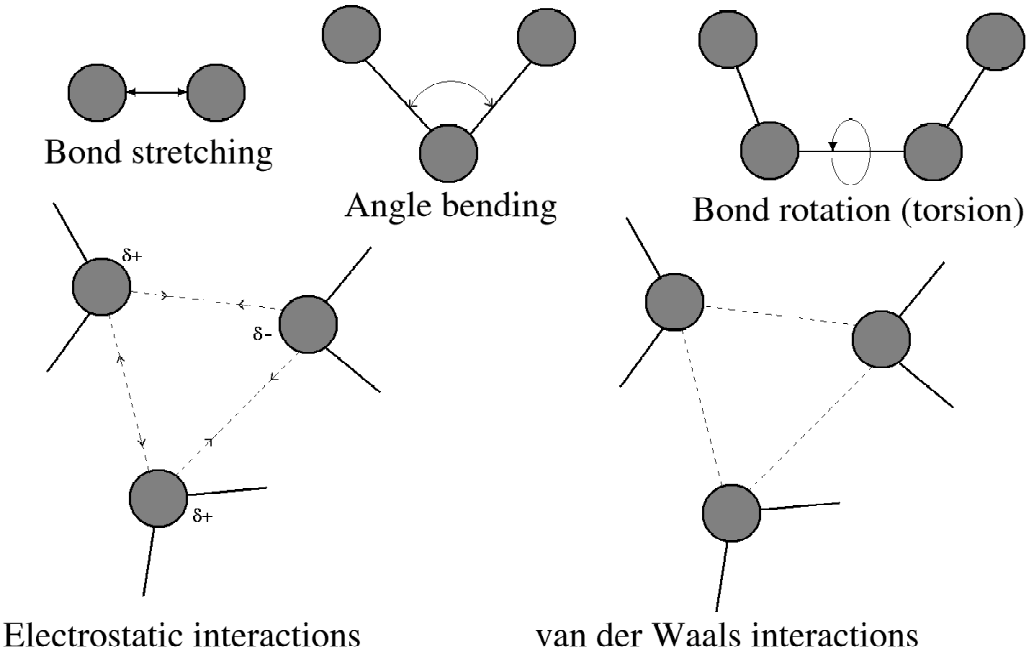


Figure 2.1. Illustration of the different energy terms used in molecular mechanics.

When using MM, the goal is often to find the optimal conformation of a molecule, by minimising the potential energy. Two commonly used minimisation techniques are the steepest descent method and the conjugate gradient method. The steepest descent method chooses the descent direction based on the energy gradient (the derivative of the energy with respect to the atomic positions) of the current step, and makes a single step in this direction. The conjugate gradient method starts along the steepest descent direction, continues along this direction until a minimum in this direction is reached, and then proceeds along a direction perpendicular (conjugate) to this direction. This is continued until a preset threshold is reached. While the steepest descent method and the conjugate gradient method use only the first derivative of the energy with respect to the atomic positions, the Newton-Raphson method also includes the second derivative. As calculation of the inverse of the second derivative leads to a high computational cost for large molecular systems, the Quasi-Newton and the Truncated Newton methods have been developed, giving approximate solutions to the problem (Jensen, 1999; Leach, 2001).

A variety of different MM force fields exist (Cramer, 2002). In this work, MMFF94 (Halgren, 1996), AMBER94 (Weiner *et al.*, 1984) and CHARMM (CHARMm22 and CHARMM27) (MacKerell *et al.*, 1998 a, b) have been used. The MMFF and CHARMm force fields have parameters for both small-molecular organic compounds and biomolecules, while AMBER and CHARMM are most suitable for use with biomolecules. CHARMm is an extended, commercial version of CHARMM (available through Accelrys Inc.).

Use of MM to predict the 3D structure of macromolecules like proteins is very time consuming, and although widely used, usually does not give reliable results within reasonable time (Baker and Sali, 2001; Leach, 2001). More approximate methods, such as homology modelling, are more suitable for this purpose. More detailed descriptions of molecular modelling can be found elsewhere (Rappé and Casewit, 1997; Jensen, 1999; Leach, 2001; Cramer, 2002; Forster, 2002).

2.1.2 Conformational searching

Because of the complexity of molecular structures, there may be more than one combination of atomic positions that give a minimum on the potential energy surface. Hence, several local minima exist in addition to the global minimum having the lowest potential energy. In particular, simulations on protein structures are multiple-minimum problems. Minimisation techniques can only guide the molecule from the starting conformation to the closest minimum on the potential energy surface. This is often not the global energy minimum. In addition to assessing discrete molecular structures, MM can be used to systematically determine the number of minima and the energetic differences between these minima. Since the global minimum problem has not yet been solved analytically, many conformational searching techniques have been developed, for example grid searching, molecular dynamics (MD), Monte Carlo (MC) simulation, simulated annealing and Tabu search (Allen and Tildesley, 1989; Rappé and Casewit, 1997; Leach, 2001; Frenkel and Smit, 2002).

Grid search methods vary each of several geometric variables in a molecule by some increment, while keeping the remaining variables fixed. Without a procedure to select in advance those conformations that are likely to have relatively low energy, this method becomes computationally expensive. In MD, successive configurations of the system under consideration are generated by integrating Newton's law of motion. This results in a trajectory that specifies how the positions and velocities of the particles in the system vary with time. In a Monte Carlo simulation, the statistical mechanical behaviour of a molecule is simulated by making random changes to the system, such as random changes in dihedral angles. The energy of a trial conformation is calculated and the changes are accepted if the energy has decreased or meets the requirement of a particular algorithm, e.g. the Metropolis criterion (Metropolis *et al.*, 1953). According to the Metropolis criterion, changes that decrease the energy of the system are always accepted, while changes that

increase the energy are accepted according to a probability distribution. Simulated annealing mimics the process of slowly reducing the temperature of a substance until it reaches thermal equilibrium. Simulated annealing is often coupled with an MC simulation.

A procedure used in this work, Tabu search (Glover and Laguna, 1993; Baxter *et al.*, 1998), is a stochastic searching algorithm, used e.g. in conformational searching. From the current conformation, a specified number of new conformations are constructed by adding random numbers to all coordinates in the current conformation. The new conformations are ranked according to an energy function. Tabu search maintains a list of previously visited conformations. These conformations are forbidden (tabu) for future moves. A new conformation is compared to the conformations in the list by calculating the root mean square deviation (RMSD) between the Cartesian coordinates of the new conformation and those of every entry in the list. If the RMSD value is below a specified value, the conformations are considered to be the same, and the move is rejected (tabu). The highest ranked conformation is always accepted (even if it is tabu) if the energy is lower than the lowest energy obtained so far in the search. Otherwise, the algorithm chooses the best non-tabu conformation. When a new current solution has been found, a new set of coordinate transformations is carried out, and the search procedure continues with the next iteration.

2.1.3 Protein structure prediction

There are four levels of protein structure: primary, secondary, tertiary and quaternary structure. The primary structure consists of the amino acid sequence, the secondary structure is built up of e.g. α-helices and β-sheets, while the tertiary structure is determined by how the different elements of secondary structure are folded. Protein quaternary structure refers to the spatial relation between different domains of tertiary structure. While α-helices and β-sheets are stabilised mostly by hydrogen bonds, protein tertiary structure is held together primarily by hydrophobic interactions (Branden and Tooze, 1999). This hydrophobic core is often well conserved within a protein family (Lesk and Chothia, 1980; Branden and Tooze, 1999). Creation of this hydrophobic core and a hydrophilic surface by packing the hydrophobic side-chains into the interior is the main driving force for folding of water-soluble globular proteins. Since the main-chain is hydrophilic, formation of α-helices and β-sheets is necessary to create the hydrophobic core. Formation of hydrogen bonds "neutralises" the polar groups of the main-chain (Branden and Tooze, 1999). Folding of a protein into its native structure is a complex process, which is not yet fully understood. The existence of chaperone proteins that assist in the folding process further complicates the problem (Creighton, 1993; Branden and Tooze, 1999). Much research is devoted to prediction of protein secondary structure, and a variety of methods exist for this purpose (Leach, 2001). However, in the following, the terms "protein structure prediction" and "3D structure prediction" refer to prediction of protein tertiary structure.

The conformational space of a macromolecule, such as a protein, is very complex, containing a large number of local energy minima separated by high free energy barriers. The conventional search methods such as molecular dynamics may sample only a small part of the conformational space due to their difficulties of overcoming high-energy barriers (Tappura *et al.*, 2000). Hence, such methods are not suitable for model building without additional information or constraints. A variety of other *ab initio* or first-principles methods for protein structure prediction also exist, such as lattice models (Chan and Dill, 1993), where the protein is modelled as a sequence of hydrophobic and hydrophilic monomers, and the energy of a conformation is calculated by summing interactions between pairs of monomers that occupy adjacent lattice sites but are not covalently bonded. Other methods use knowledge-based rules for packing of different secondary structure elements to arrange α-helices and β-sheets into a low-energy structure (Cohen *et al.*, 1982).

At the present time, homology modelling (also called comparative modelling) (Bajorath *et al.*, 1993; Sánchez and Sali, 1997; Marti-Renom *et al.*, 2000) is usually the fastest way to generate an

approximate model of a protein structure when 3D structures of related proteins are available to be used as templates. Homology modelling is based on the observation that proteins having related primary structure (proteins that have diverged from a common ancestor protein during evolution) share segments of similar conformation. It is assumed that if the amino acid sequences are closely related, then the 3D structure of a protein can be predicted from the known 3D structures of other proteins within the same family. However, even for closely related proteins there are unique regions, which can differ significantly both in sequence and conformation (Tappura *et al.*, 2000). Such regions are generally surface loops connecting the regular secondary structures. Unfortunately, such surface loops are often important for biological activity and diversity of a binding site. Protein structure prediction and homology modelling have recently been reviewed (Al-Lazikani *et al.*, 2001; Baker and Sali, 2001; Schonbrun *et al.*, 2002).

The homology modelling methodology can be divided into four main steps:

1. Identification of one or more suitable template structures.
2. Generation of an alignment between the target and template amino acid sequences.
3. Generation of a structural model based on the sequence alignment.
4. Validation of the model.

Different homology modelling methods are described in detail in Chapter 2.3.1.

The quality of homology models is highly dependent on the choice of template structures. A protein structure can provide a close general model for other proteins with which its sequence similarity is higher than 50% (Chothia and Lesk, 1986). If the sequence similarity drops to 20%, there will be large structural differences. It has been indicated that in general, a sequence similarity of about 45-60% is needed for the homology models to be used for virtual screening (Shoichet *et al.*, 2002). However, the active sites of distantly related proteins can have very similar geometries (Lesk and Chothia, 1980; Chothia and Lesk, 1982). A weakness of using structures predicted by homology modelling as basis for the design of selective drugs is that to achieve selectivity one has to utilise variable regions of the proteins. These are the regions predicted with the lowest reliability by homology modelling techniques (Read *et al.*, 1984).

In domain modelling, the positions of any atoms forming an interface to a missing domain should be fixed during energy minimisation. Free movement in these regions can lead to side-chain conformations that are preferable energetically, but not possible in the real protein structure due to interactions with the missing parts of the protein.

In cases where no template structure of sufficient sequence similarity exists, a method called fold recognition (or “threading”) can be used (Torda, 1997). Fold recognition is based on knowledge that most proteins fold into one of a limited number of stable folds. In fold recognition, databases of known protein folds are searched to find the fold that the query sequence is most likely to adopt, using a “pseudo-energy” function.

2.1.4 Binding free energy estimation

To evaluate how strongly a ligand, such as a drug molecule, binds to its receptor, the binding free energy (the binding affinity) is estimated. When activity is directly associated with ligand binding, this provides a measure of how effective the ligand is in affecting the activity of the receptor. This is important in drug design, where the goal is to identify or construct a ligand with as high affinity towards the target receptor as possible.

Gibbs free energy change (at a temperature T) associated with binding of a ligand to a receptor is defined as

$$\Delta G_{binding} = G_{complex} - \left(G_{receptor} + G_{ligand}\right) = \Delta H_{binding} - T\Delta S_{binding} = -RT \ln K_i \quad (2.2)$$

where $G_{complex}$ is the free energy of the ligand-receptor complex and $G_{receptor}$ and G_{ligand} refer to the free energy of the ligand and the receptor, respectively, prior to binding. The enthalpic contributions ($\Delta H_{binding}$) result from interatomic forces such as electrostatic and van der Waals forces, while $\Delta S_{binding}$ represents the entropy. K_i is the binding constant, and R is the gas constant. Numerous methods exist for estimation of the binding free energy (Leach, 2001; Cramer, 2002).

The free energy contains a solvation term (G_{solv}), representing the free energy change accompanying the transfer of a molecule from vacuum to solvent, a Coulombic term (G_{coul}), representing attractive and repulsive forces between charged particles in the system and a van der Waals term (G_{vdW_solute}) representing repulsive and dispersive forces between the particles in the solute. The free energy may be calculated as

$$G=G_{solv}+G_{coul}+G_{vdW\_solute}=G_{solv\_ele}+G_{solv\_vdW}+G_{cav}+G_{coul}+G_{vdW\_solute}=G_{ele}+G_{surf}+G_{vdW\_solute} \quad (2.3)$$

where G_{solv_ele} represents the electrostatic contribution to the solvation free energy, G_{solv_vdW} is the van der Waals interaction between the solute and the solvent and G_{cav} is the free energy required to form the cavity within the solvent containing the solute molecule. G_{ele} is the total electrostatic contribution. G_{solv_vdW} and G_{cav} can be computed together using the total solvent-accessible surface area (A_{acc}):

$$G_{surf} = G_{solv\_vdW} + G_{cav} = \gamma A_{acc} + b \quad (2.4)$$

where γ and b are constants. For systems with hydrogen bonding between the solute and the solvent, an additional hydrogen-bonding term may be added to Equation 2.3 (Leach, 2001). A term representing the internal energy of the solute molecule(s) may also be added.

For binding of a ligand to a receptor to be favourable, the energy of the solvated ligand-receptor complex has to be lower than the energy of the solvated ligand and receptor prior to binding. Hence, the solvation energy is an important factor in ligand binding. Several different approaches exist for estimating the solvation free energy (Leach, 2001). One approach to account for solvation effects is addition of explicit water molecules to the system prior to MM or MD calculations. This approach is time consuming, and not applicable to large molecular systems. An alternative is to use continuum solvation models, which treat the solvent as a continuous medium having the average properties of the real solvent. The solvent surrounds the solute, beginning at or near its van der Waals surface (Qiu *et al.*, 1997).

The electric field at a given point in space is the gradient of the electrostatic potential $\phi(\mathbf{r})$ at that point, and the work required to create the charge distribution can be determined from the interaction of the solute charge density $\rho(\mathbf{r})$ with the electrostatic potential from the surroundings according to Equation 2.5 (Cramer, 2002).

$$G = -\frac{1}{2}\int \rho(\mathbf{r})\phi(\mathbf{r})\, d\mathbf{r} \quad (2.5)$$

An approach to estimate the electrostatic contribution to the solvation energy that has been particularly useful for biological macromolecules such as proteins is based on solving the Poisson equation (Leach, 2001). The Poisson equation relates the variation in the potential ϕ within a medium of uniform dielectric constant ε to the charge density ρ:

$$\nabla^2\phi(\mathbf{r}) = -\frac{4\pi\rho(\mathbf{r})}{\varepsilon} \quad (2.6)$$

∇^2 is defined by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.7}$$

The Poisson equation is valid under conditions of zero ionic strength. If mobile electrolytes are present in the solvent, the Poisson-Boltzmann (PB) equation applies instead. The Poisson equation can be considered a special case of the PB equation (Cramer, 2002). In the Poisson-Boltzmann surface area (PB/SA) method, the Coulombic term is calculated explicitly:

$$G_{ele} = G_{solv\_ele} + G_{coul} \tag{2.8}$$

G_{solv_ele} is obtained by solving the Poisson-Boltzmann equation (or the Poisson equation) and G_{coul} is the Coulombic energy associated with ligand binding to the protein. For a system of particles with interparticle distances r_{ij} and charges q_i, the Coulombic energy is calculated as

$$G_{coul} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{\varepsilon r_{ij}} \tag{2.9}$$

where ε is 1 in gasphase and approximately 78 in water.

In order to solve the Poisson equation, numerical methods are needed, since the Poisson equation has not yet been solved analytically for an arbitrary shape of the molecule. An alternative is to use the Generalised Born (GB) equation, which is an approximation to the Poisson equation that can be solved analytically (Cramer, 2002). In the Generalised Born surface area (GB/SA) method, G_{ele} is estimated as

$$G_{ele} = G_{solv\_ele} + G_{coul} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right) \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{f(r_{ij}, a_{ij})} \tag{2.10}$$

The following form for the function $f(r_{ij}, a_{ij})$ has been proposed (Qiu *et al.*, 1997):

$$f(r_{ij}, a_{ij}) = \sqrt{(r_{ij}^2 + a_{ij}^2 e^{-D_{ij}})} \tag{2.11}$$

where $a_{ij} = \sqrt{(a_i a_j)}$ and $D_{ij} = r_{ij}^2 / (2a_{ij})^2$, and a_i represents the radii of the particles. Hence, for large interparticle distances r_{ij}, $f(r_{ij}, a_{ij})$ is approximately equal to r_{ij}.

2.1.5 Surface area calculations

As discussed in the previous section, calculated surface areas can e.g. be used to estimate non-covalent interactions between molecules. The van der Waals surface areas, the molecular surface areas and the solvent accessible surface areas are commonly used for this purpose (Leach, 2001). The van der Waals surface is constructed from the overlapping van der Waals spheres of the atoms (Figure 2.2).

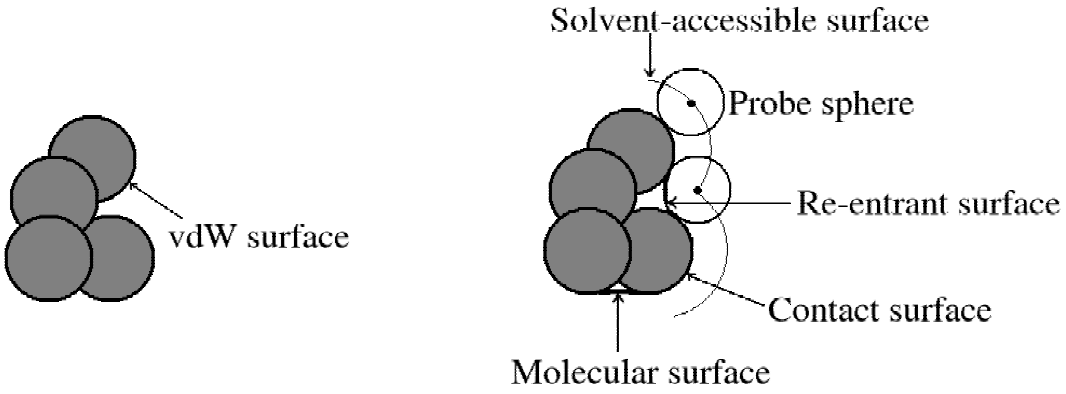


Figure 2.2. Calculation of surface areas of molecules. The molecular surface consists of the contact surface and the re-entrant surface.

Surface areas are usually calculated by rolling a probe of a specified radius over the van der Waals surface of the given atom or molecule, tracing the centre of the probe. The molecular surface is traced out by the inward-facing part of the probe sphere, and contains two different types of surface element, the contact surface and the re-entrant surface. The contact surface corresponds to the regions where the probe is actually in contact with the van der Waals surface of the molecule. The re-entrant surface regions occur where there are crevices that are too narrow for the probe molecule to penetrate. The molecular surface is usually estimated using a water molecule, represented by a sphere of radius 1.4 Å, as the probe. The solvent accessible surface area is the surface that is traced by the centre of the probe molecule. The centre of the probe can thus be placed at any point on the accessible surface and not penetrate the van der Waals spheres of any of the atoms in the molecule. Several algorithms for calculating molecular and accessible surface areas have been published (Connolly, 1983; Richmond, 1984; le Grand and Merz Jr., 1993).

2.2 Multivariate regression

Regression analysis is the process of relating a set of independent variables (called the **X**-matrix) to one or more dependent variables, or response variables (the **Y**-matrix), through a matrix of regression coefficients (**B)**:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F} \tag{2.12}$$

F is the residual matrix, that is, the part of **Y** that is not described by the regression model. The purpose of the regression analysis is to train a regression model that can be used to predict the response for new samples. In Partial Least Squares (PLS) regression, a decomposition of **X** and **Y** into a latent variable space is carried out, where the purpose is to maximise the covariance between **X** and **Y** (Martens and Martens, 2000; Høy, 2002). This is in contrast to Ordinary Least Squares (OLS) regression, where the correlation between **X** and **Y** is maximised. PLS is superior to OLS for example in cases where the X-variables are correlated. In PLS, **X** and **Y** are related through a common score matrix (**T**), as shown in Figure 2.3 and Equations 2.13 and 2.14.

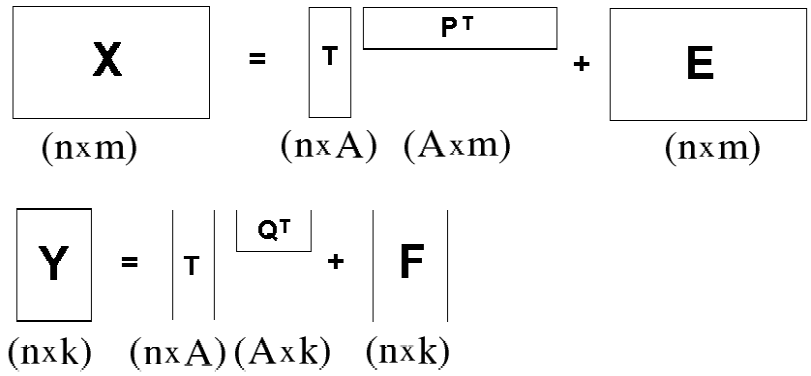


Figure 2.3. Decomposition of **X** and **Y** into latent variable space in PLS regression. n represents the number of samples, m and k are the number of X- and Y-variables, respectively, and A is the number of principal components (PCs).

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.13}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F} \tag{2.14}$$

The matrices **P** and **Q** represent the X- and Y-loadings, respectively, while **E** and **F** are the residual matrices. Columns of **T**, **P** and **Q** corresponding to insignificant PCs are not used. Hence, **E** and **F** depend on the number of PCs used. The number of significant PCs is usually chosen based on the explained Y-variation from a validation of the regression model. A commonly used technique is leave-one-out cross-validation, where each sample is kept out of the regression analysis in turn, and the response is predicted using the remaining samples. This gives a measure of the predictive power of the regression model. The loading weights (**W**) are defined by

$$\mathbf{T} = \mathbf{X}\mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \tag{2.15}$$

X-variables for new samples can be used to predict the response (Y) using the regression coefficients from the PLS regression:

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{Q}^T \tag{2.16}$$

$$\mathbf{Y}_{predicted} = \mathbf{X}_{new}\mathbf{B} \tag{2.17}$$

More complete revisions of PLS regression can be found elsewhere (Manne, 1987; Høy, 2002).

2.3 Rational drug design methods

2.3.1 Homology modelling methods

Homology modelling methods can be divided into three main groups: Rigid body superposition methods, methods based on distance geometry and segment matching methods. In rigid body superposition, a model is constructed from a few core sections defined by the average of the Cα atoms in the conserved regions. Distance geometry uses spatial restraints obtained from the alignment, while in segment matching, a database of short segments of protein structure is used, together with energy or geometry rules. Examples of available homology modelling programs

include SwissModel (Peitsch, 1995; Peitsch, 1996; Guex and Peitsch, 1997; Guex *et al*., 1999; Schwede *et al.*, 2003), WHAT IF (Vriend, 1990), MODELLER (Sali and Blundell, 1993; Fiser *et al*., 2000; Marti-Renom *et al*., 2000) and LOOK (Levitt, 1992).

SwissModel is a popular implementation of the rigid body approach. A model framework is first generated by ProModII (Peitsch, 1996), based on the topological arrangement of corresponding atoms to the given templates. The backbone is rebuilt based on the positions of Cα atoms, using a library of backbone elements derived from high quality X-ray structures. Incomplete loops and incomplete or lacking side-chains are rebuilt prior to an energy minimisation with molecular mechanics.

The homology modelling procedure in WHAT IF starts with copying the backbone of the template structure. The side-chains of the different residues are then placed in order according to the narrowness of the position-specific rotamer distribution (Chinea *et al.*, 1995). The side-chains of the residues having the narrowest rotamer distribution are placed first. The rotamer distribution is determined by extracting from a database of non-redundant protein structures all suitable fragments of five or seven residues. Suitable fragments are those that have a local backbone conformation similar to the one around the evaluated position, and have the same residue type at the actual position. Rotamers are rejected if they lead to severe van der Waals clashes when placed in the model.

In MODELLER, restraints on distances and dihedral angles are generated based on the target-template alignment. Corresponding distances and angles between aligned residues in the template and the target structures are assumed to be similar. Restraints on bond lengths, bond angles, dihedral angles and non-bonded atom-atom contacts are derived from the CHARMM force field and from statistical analysis of the relationships between Cα atoms, solvent accessibilities and side-chain torsion angles in known protein structures. The restraints are expressed as probability density functions (pdfs). These pdfs are combined to give a molecular function, which is optimised by combining energy minimisation with molecular dynamics and simulated annealing.

LOOK uses Segment Match Modelling (SegMod) to generate homology models by fragment-based assembly (Kolodny *et al.*, 2002). SegMod uses a fragment-matching algorithm to find the appropriate structural segments derived from known 3D structures. Both backbone and side-chain information from the fragments are utilised to obtain the model, which is energy minimised using molecular mechanics. SegMod models insertions and deletions by searching for compatible fragments.

2.3.2 Methods for verification of the accuracy of protein structure models

An inaccurate protein structure model may be misleading, and relatively small structural errors may lead to large errors in e.g. binding energy calculations. Hence, it is important to be able to predict the reliability of protein structure models prior to applying them in e.g. drug design. The homology model accuracies are comparable for most modelling methods when the methods are used optimally (Koehl and Levitt, 1999). However, automatic methods will not always find the optimal alignments or loop predictions, especially when the sequence identity falls below 40% (Marti-Renom *et al.*, 2000; Qian and Goldstein, 2002). Misalignments and errors in the loop modelling are the largest sources of errors in comparative modelling (Fiser *et al.*, 2000; Marti-Renom *et al.,* 2000). Several methods exist for prediction of the reliability of sequence alignments (Cline *et al.*, 2002; Tress *et al.*, 2003)

Models of protein 3D structures can be evaluated according to a variety of criteria, such as stereochemistry, bond lengths, bond angles, torsion angles, packing, formation of a hydrophobic core, residue and atomic solvent accessibilities, spatial distribution of charged groups, distribution of atom-atom distances, atomic volumes and main-chain hydrogen bonding. Large deviations from the most likely values have been interpreted as indicators of errors in the model structure. Examples

of such methods include PROCHECK (Laskowski *et al.*, 1993), AQUA (Laskowski *et al.*, 1996), SQUID (Oldfield, 1992) and WHATCHECK (Hooft *et al.*, 1996). Methods based on 3D profiles and statistical potentials of mean force also exist, that take many of these criteria into account implicitly. These methods evaluate the environment of each residue as seen in the model, compared to the expected environment as observed in experimental structures. Examples of programs utilising such methods include VERIFY3D (Lüthy *et al.*, 1992), PROSAII (Sippl, 1993), HARMONY (Topham *et al.*, 1994) and ANOLEA (Melo and Feytmans, 1998). WHATCHECK reports have also been used in the refinement of homology models with restricted molecular dynamics (Flohil *et al.*, 2002). For each residue, the probability that it is modelled correctly is determined. During the MD simulation, correctly modelled residues are restrained, while those likely to be wrongly positioned are allowed to move unrestricted. Recently, analysis of backbone deviations between pairs of homologous proteins was used to predict local backbone structural deviation in homology models (Cardozo *et al.*, 2000). This method may be used to evaluate the utility of a preliminary homology model for e.g. drug design or to provide an improved starting point for loop prediction.

The accuracy of protein structure models can also be evaluated by comparison to experimental structures of the targets (Venclovas *et al.*, 1997; Cristobal *et al.*, 2001; Moult *et al.*, 2001; Moult *et al.*, 2003). A common method is to use RMSD values between the positions of corresponding atoms in the two protein 3D structures. However, the geometric measures only provide meaningful results when the entire extent of the proteins is comparable. For example, a set of partially correct structures cannot be ranked because the incorrect portions will dominate the RMSD value. When restricting the comparison to certain parts of the structure, the definition of relevant parts is also not always obvious. An alternative is to compare the surface areas of residue contacts in the protein structures. This procedure does not require a superpositioning of the structures that are being compared.

When estimations of surface areas of residue contacts are used to evaluate the accuracy of a protein structure model, the contact areas between all pairs of residues in the model structure are calculated and compared to the results obtained for a reference structure. How well the residue-contacts in the model correspond to the same contacts in the reference structure is then used as a measure of the model accuracy. The contact area A_{ij} between residues i and j of a protein is calculated by identifying the part of the surface area of residue i that is occluded by van der Waals surfaces of atoms of residue j. The matrix containing A_{ij} for each pair of residues in a protein structure is referred to as the contact area matrix. When two protein structures are compared, the difference between the contact area matrices for the two structures is calculated. The elements in the resulting matrix are negative for incorrectly occurring and overestimated contacts, zero for correct contacts and non-contacting residue pairs, and positive for underestimated or missing contacts in the model structure. In the following, this matrix will be referred to as the inter-residue contact area error matrix. This contact area error matrix can be summed over all elements to give a single value representing the model error, the Contact Area Difference (CAD) number. The CAD number for a reference structure R and a model structure M is given by Equation 2.18 (Abagyan and Totrov, 1997).

$$CAD = \sum_{i,j} \left| \left(A_{ij}^{R} - A_{ij}^{M} \right) \right| \tag{2.18}$$

The CAD number can be normalised to make it independent of e.g. protein size, shape and amino acid content (Abagyan and Totrov, 1997).

Recently, a new surface area based comparison method has been developed (Liu *et al.*, in preparation). This method is similar to the CAD number calculation described above (Abagyan and Totrov, 1997), but differs in both calculational details and in the normalisation of the CAD number. Here, the normalised CAD number is calculated as

$$CAD_{norm} = \frac{\sum_{i,j} \left| \left(A_{ij}^{R} - A_{ij}^{M} \right) \right|}{\frac{N}{2} \left(A_{ii}^{R} + A_{jj}^{R} \right)} \tag{2.19}$$

where N is the number of residues considered. The surface areas are calculated using a Boolean logic based algorithm (le Grand and Merz Jr., 1993). Analysis of residue-residue contacts has been used to evaluate structure predictions (Göbel *et al.*, 1994), and the conservation of side-chain interactions in homologous proteins (Russell and Barton, 1994). Contact-based measures have also been applied to study protein folding using simplified protein descriptions (Gou and Thirumalai, 1995).

2.3.3 Methods for mapping protein binding site properties

A variety of methods exist for localisation of protein binding sites by detecting cavities in the structures. Examples include Putative Active Sites with Spheres (PASS) (Brady and Stouten, 2000), Automatic PROtein Pocket Search (APROPOS) (Peters *et al.*, 1996) and CAST (Liang *et al.*, 1998). These methods use sphere-based approaches to detect grooves or pockets in the protein structure. LIGSITE (Hendlich *et al.*, 1997) recognises binding sites by evaluating the degree of surface depression burial for each point of a 3D grid surrounding the protein structure. In the present work, we focus on methods for analysis of the properties of an already known binding site, and identification and characterisation of possible interaction sites for ligands.

Numerous methods for analysis of binding site properties are available (Sotriffer and Klebe, 2002). Most of these methods identify favourable binding locations by placing atom probes, molecular fragments or small molecules at various points in the binding site and evaluating their interactions with the protein. Protein flexibility is usually not accounted for in these calculations. One class of methods is based on using a discrete 3D grid to position the probe atoms or groups within the binding site, and using an energy function to compute the interaction energies between the protein and the probes. One of the most common methods in this class is GRID (Goodford, 1985).

An alternative to the grid-based approaches is the multiple copy simultaneous search (MCSS) method (Miranker and Karplus, 1991). This method identifies favourable interaction sites in a protein cavity by placing a large number of copies of one or more probe molecules into the active site of the target protein. These probes are placed randomly around the active site atoms. The probe groups are then subjected to energy minimisation along with a molecular dynamics simulation. The receptor atoms may be kept fixed, or be subject to the average forces of the probes (Stultz and Karplus, 1999). Each probe is subject to the full force of the receptor but not forces from the other probes. Hence, interactions between the probes are not considered. Favourable interaction sites for drug candidates can then be identified based on the distribution of the different types of molecular fragments in the protein binding site.

Compared to the grid-based approaches, MCSS has the advantage that the positions of the fragments are not restricted to predefined grid points, but free to move to a more optimal location. However, the fixed grids have a major advantage with respect to comparability. Many related proteins can be superpositioned and the same grid used for all of them. Interesting differences between related binding sites can be identified by comparing energy maps. Thus, using e.g. GRID on multiple proteins can aid the development of ligands selective for a particular protein target. The data from the GRID computations can be analysed with e.g. Principal Component Analysis (PCA) (Johnson and Wichern, 1998) to find the most important structural differences to take into consideration in the design of a selective inhibitor (Pastor and Cruciani, 1995).

A useful approach for visual inspection of binding site properties is mapping of physicochemical properties such as the electrostatic potential onto molecular surface representations. The electrostatic potential is most commonly calculated by solving the Poisson-Boltzmann equation. This can be done with programs such as the University of Houston Brownian Dynamics (UHBD) program (Davis *et al.*, 1991; Madura *et al.*, 1995). An example of a method utilising this approach is GRASP (Nicholls *et al.*, 1991). GRASP contains a Poisson-Boltzmann solver in addition to visualisation routines. Methods for generation of hydrophobicity maps of protein binding sites have also been developed (Scarsi *et al.*, 1999). Here, a non-polar probe is rolled over the protein surface and the binding energy is calculated based on the van der Waals interaction and the electrostatic desolvation energy of the protein.

Rule-based or knowledge-based methods for mapping of protein binding site properties also exist. These methods use rules for preferred protein-ligand interaction patterns derived from statistical analysis of the structural data stored in databases of experimental structures of protein-ligand complexes. The program LUDI makes use of such statistical rules to calculate interaction sites suitable for hydrophobic contacts or for hydrogen bond formation (Böhm, 1992 a, b). A program called SUPERSTAR (Verdonk *et al.*, 1999; Verdonk *et al.*, 2001) identifies interaction sites in proteins based on the information stored in the database ISOSTAR (Bruno *et al.*, 1997). Here, 3D maps showing the propensities of different probes at different positions in the protein binding site are generated. Gaussian functions have been used to obtain smoother propensity maps from ISOSTAR (Nissink *et al.*, 2000).

2.3.4 Computational docking methods

Property maps of protein binding sites can be used to search databases for known drugs having properties that match the binding site properties. The hits from for example database searching can be evaluated further by molecular docking. Maps of protein binding site properties are also utilised directly in many computational docking methods. In computational docking, the ligand structure is placed in the protein binding site, and the most favourable binding conformation is sought. This is done by maximising the complementarity between a description of the protein binding site and the properties of different ligand conformations (Figure 2.4). Sometimes receptor flexibility is also taken into account in the calculations.

Docking methods use a conformational search method to optimise the bound ligand conformation, and a score function to guide the conformational search by estimating the binding affinity for the different conformations. Available search methods range from rigorous search methods such as simulated annealing (Kirkpatrick *et al.*, 1983) to faster methods such as Tabu search (Baxter *et al.*, 1998) and genetic algorithms (Terfloth and Gasteiger, 2001; Halperin *et al.*, 2002). Commonly used docking programs include DOCK (Ewing and Kuntz, 1997), AutoDock (Morris *et al.*, 1998), Molecular Operating Environment (MOE)-Dock (Hart and Read, 1992; Baxter *et al.*, 1998) and FlexX (Rarey *et al.*, 1996). Existing docking and virtual screening methods have recently been reviewed (Bajorath, 2002; Halperin *et al.*, 2002; Lyne, 2002; Taylor *et al.*, 2002; Brooijmans and Kuntz, 2003).

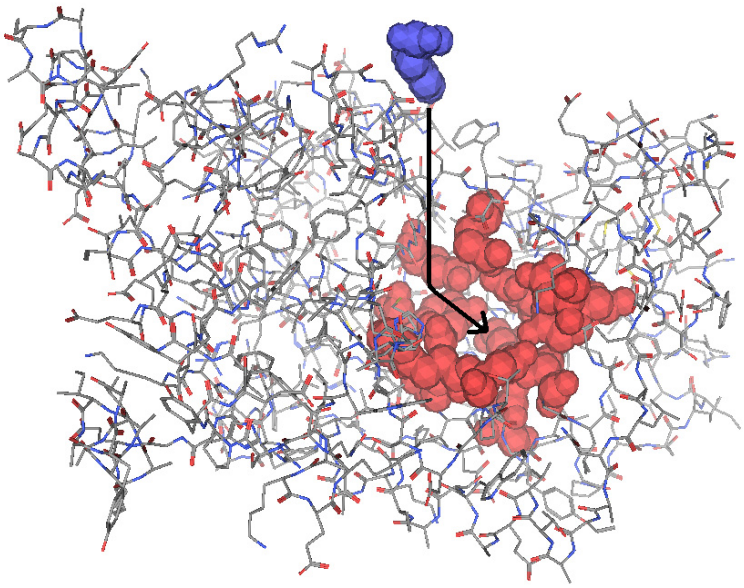


Figure 2.4. Illustration of computational docking of a ligand into a protein binding site. The ligand (blue) and the protein binding site (red) are rendered as "space filling". Based on a conformational search algorithm, a large number of ligand conformations are evaluated for binding to the receptor.

DOCK describes the ligand and the receptor cavity as two sets of spheres, and orients the ligand to maximise the complementarity between the ligand and receptor spheres. An intermolecular score is calculated based on the AMBER force field, where the receptor terms are calculated on a grid (Meng *et al.*, 1992). In DOCK 4.0, ligand flexibility is included by incorporating an intramolecular score for the ligand into the score function (Makino and Kuntz, 1997). Further extensions of DOCK have included protein flexibility using ensembles of protein structures (Knegtel *et al.*, 1997) and a GB/SA continuum model into the score function (Still *et al.*, 1990; Zou *et al.*, 1999; Liu *et al.*, 2004).

Early implementations of AutoDock used Monte Carlo simulated annealing with a grid-based evaluation of the energy based on the AMBER force field, to dock flexible ligands into the binding pocket of a rigid receptor (Goodsell and Olson, 1990; Morris *et al.*, 1996). A more recent version uses a genetic algorithm combined with an energy minimisation for the conformational search (Morris *et al.*, 1998). The score function includes five terms: van der Waals energy, desolvation energy, a directional hydrogen bonding term, a term representing the Coulombic electrostatic energy and a term proportional to the number of sp^3 bonds in the ligand, representing unfavourable entropy of ligand binding due to the restriction of conformational degrees of freedom. Protein and ligand parameters are taken from the AMBER force field.

With MOE-Dock, two different docking procedures are possible; one using Monte Carlo simulated annealing for the conformational search (Hart and Read, 1992), and one using Tabu search (Baxter *et al.*, 1998). The score function is a sum of the electrostatic and the van der Waals interaction energy between the ligand and the target protein, and the intramolecular energy of the ligand. To calculate the interaction energies, MOE uses a grid-based method where the interaction energy is calculated using electrostatic and van der Waals fields that have been sampled on a grid overlaying the docking box inside which the ligand is allowed to move. Hence, this grid-based method calculates the potential energy grids only once, at the beginning of the docking procedure. The energy fields are interpolated at the atom positions by tri-linear interpolation. The van der Waals parameters are taken from the currently active force field (several different force fields are possible, for example MMFF and AMBER), and the electrostatic energy is calculated in a Coulombic manner (MOE, 2002).

FlexX is a fragment-based docking method that builds up the ligand using an incremental construction algorithm (Rarey *et al.*, 1996). Following an initial base fragment selection, different ligand conformations are formed based on the MIMUMBA torsion angle database (Klebe and Mietzner, 1994). Intramolecular and intermolecular overlaps are removed, and the conformations are ranked using an empirical score function that accounts for hydrogen bonds, ionic interactions, the lipophilic protein-ligand contact surface and the number of rotatable bonds in the ligand (Böhm, 1994). A recent version of FlexX includes explicit water molecules into the binding site using pre-computed water positions (Rarey *et al.*, 1999). A version of FlexX suited for combinatorial library docking, FlexX^c, has also been developed (Rarey and Lengauer, 2000).

Recently, a docking program called EasyDock has been developed, that makes use of quantum stochastic tunnelling for conformational searching (Todorov *et al.*, 2003; Mancera *et al.*, 2004). The method combines the use of multiple ligand copies with a non-linear transformation of the potential energy surface that allows for the positions of the local minima to be retained while the sizes of the transition barriers connecting them are significantly reduced. This reduces the probability of getting trapped in local minima with high-energy transition-state barriers, a problem associated with e.g. simulated annealing.

SLIDE (Screening for Ligands by Induced-fit Docking, Efficiently), is a fast docking method that is suitable for virtual screening (Schnecke and Kuhn, 1999 a, b; Schnecke and Kuhn, 2000). Templates with hydrogen bonding and hydrophobic interaction points for the protein binding site and for the ligand candidates are first generated, where an interaction point is a hydrogen bond donor or acceptor, or a hydrophobic ring centre. The best matches between triplets of interaction points on the ligands and triplets of receptor template interaction points are calculated based on chemical properties and geometry. Template triangles for triplets of ligand interaction points are docked into the binding site by least squares fit of the ligand triangles onto the receptor template triangles. Rigid anchor fragments for the ligands are then generated based on matched interaction point triangles, flexible bonds in the ligands are identified, and collisions between ligand anchor fragments and the protein are resolved by iterative ligand translations. Side-chain collisions are resolved by directed rotations. The protein-ligand complexes are then scored based on the number of intermolecular hydrogen bonds and hydrophobic complementarity. SPECITOPE ("Specific Epitope") is an earlier version of this program (Schnecke *et al.*, 1998).

Recently, new docking methods especially suited for use with homology modelled protein structures have been developed. Gaussian functions have been used to represent the physicochemical properties of the receptor and the ligand in computational docking (Schafferhans and Klebe, 2001). This method optimises the overlap between the functional description of the receptor binding site and the ligands. Ligand information is also incorporated into the protein structure modelling procedure (Schafferhans and Klebe, 2001; Evers *et al.*, 2003). A gaussian-based docking method that is meant to act as a filter to reduce the search space for other docking methods has recently been developed (McGann *et al.*, 2003). This method accounts only for shape, and minimises steric clashes between the receptor and ligand atoms. Another docking method suitable for homology models uses a discretisation of the structural models, together with an averaging of the structural details and a smoothing of the potential energy surface to compensate for structural errors (Wojciechowski and Skolnick, 2002). Both steric and chemical complementarity between the ligand and the receptor is sought using a grid-based search.

To increase computational efficiency, protein flexibility has traditionally been ignored in docking calculations. This is a severe approximation, since it is well known that in many cases the ligand induces conformational changes in the protein structure, a process called induced fit (Ishima and Torchia, 2000; Ma *et al.*, 2002; Teague, 2003). When using protein structure models built by homology modelling, it is especially important to allow for protein flexibility, since this can reduce the impact of small structural errors. Ligands present in the X-ray structures used as templates in the homology modelling may also have induced conformational changes in the protein. Using a fixed protein structure might thus hinder identification of the correct binding modes for other ligands.

Examples of methods that use side-chain flexibility include SLIDE (Schnecke and Kuhn, 1999 a, b; Schnecke and Kuhn, 2000), the method reported by Leach (1994), which includes side-chain flexibility using information from analysis of high-resolution protein structures, and the "Mining Minima Optimiser" method (Kairys and Gilson, 2002), where rotatable bonds in selected protein side-chains can be treated as continuous degrees of freedom during the docking procedure. An algorithm for identifying regions where conformational adaptation to a ligand is likely to occur has also been developed (Anderson *et al.*, 2001). During the docking simulations the side-chains of these residues are allowed to move. Rotamer libraries have also been used to include side-chain flexibility (Schaffer and Verkhivker, 1998).

One approach to include protein flexibility in docking calculations is to use soft docking. In soft docking, the high energy penalty for overlap between ligand and receptor atoms is relaxed. This can be done by reducing the van der Waals contributions to the total energy score. An example of a soft docking method is the method developed by Jiang *et al.* (1991), where the molecular volumes and surfaces are represented as cubes that are first matched geometrically and then scored according to the favourable energetic interactions between the buried surface areas. The majority of the methods that take protein backbone flexibility into account utilise multiple protein structure models in the calculations. In the "Relaxed Complex Method" (Lin *et al.*, 2002; Lin *et al.*, 2003) a long molecular dynamics simulation of the unliganded receptor is carried out, followed by a rapid docking of candidate ligands to a large ensemble of the receptor's MD conformations. The use of statistical analysis of conformational samples from short-run protein molecular dynamics has also been combined with grid-based docking by generation of a composite interaction weight-averaged grid (Broughton, 2000). Experimental protein structures have also been used to generate combined interaction grids by averaging with respect to energy and geometry (Knegtel *et al.*, 1997). The FlexE approach (Claussen *et al.,* 2001), a variant of FlexX, is based on a united protein description generated from an ensemble of protein structures. Discrete alternative conformations are explicitly taken into account for varying parts of the protein. These conformations can be combinatorially joined to create new protein structures. Structural water heterogeneity has also been incorporated into docking simulations, in addition to protein flexibility, using an ensemble of protein structures (Österberg *et al.*, 2002). Recently, a hybrid approach where the first component is ligand docking to a rigid receptor, and the second step is an MC simulation including the GB/SA continuum solvent model has been developed (Taylor *et al.*, 2003). A rotamer library is also used to direct some of the protein side-chain movements along with large dihedral moves, and a softening function is used for the non-bonded force field terms.

A novel algorithm called IFREDA (Internal coordinate mechanics (ICM)-flexible receptor docking algorithm) generates an ensemble of receptor conformations by performing flexible ligand docking of selected known binders to a flexible receptor (Cavasotto and Abagyan, 2004). This ensemble is then used to perform flexible ligand-rigid receptor docking. Docking of known binders has also been used to select a minimal subset of receptor conformations that provides a strong correlation between the experimental binding affinities and the docking scores (Yoon and Welsh, 2004). This subset is then used for multiple-conformation docking.

2.3.5 Score functions for computational docking

As previously mentioned, docking methods use score functions to guide the conformational search and to estimate the binding energies between the receptor and the ligand. Existing score functions can be divided into three main categories: force field based, empirical and knowledge-based score functions. The use of score functions in drug design has recently been reviewed (Böhm and Stahl, 2002).

Force field based scoring methods estimate the binding affinity using non-bonded energies of molecular mechanics force fields. Force field based score functions are often slow and sensitive to

errors in the protein structure models, partial charges and protonation states. Examples of force field based methods include the score function implemented in the AutoDock program (Morris *et al.*, 1998) which utilises parameters from the AMBER force field and MM PB/SA (Massova and Kollman, 1999) which includes a solvation term calculated by the Poisson-Boltzmann equation (Honig and Nicholls, 1995) in addition to the electrostatic interactions. The OWFEG (one window free energy grid) method (Pearlman and Charifson, 2001) is an approximation to the computationally expensive free energy perturbation method (Meirovitch, 1998). In OWFEG, a molecular dynamics simulation is carried out with the ligand-free, solvated receptor site. Solvent effects are represented explicitly. The energetic effects of probe atoms on a regular grid are collected and averaged during the simulation. Three simulations are run with three different probes: a neutral atom, a negatively charged and a positively charged atom. This results in three energy fields, containing information about the score contributions of neutral, positively and negatively charged ligand atoms located in various positions of the receptor site. The advantages of the OWFEG method are the consideration of entropic and solvent effects, and the inclusion of some protein flexibility in the simulations. This is achieved by allowing weakly restrained motion of the region of the protein near the active site and free movement of solvation water molecules.

Empirical score functions are generally faster than force field based methods. It is assumed that the binding free energy can be interpreted as a weighted sum of localised interaction terms, representing hydrogen bonds, ionic interactions, hydrophobic interactions, entropy change associated with binding, etc. In addition, penalty functions for e.g. steric clashes can be added. The interaction terms are usually calculated using experimental 3D structures of receptor-ligand complexes, and the weights are estimated by multiple linear regression using experimental binding affinities. This makes the empirical score functions very dependent on the set of experimental structures used to train the parameters. Usually, between 50 and 100 complexes are used to train the score functions, but recently it was shown that more than 100 complexes are needed for convergence (Wang *et al.*, 1998). A disadvantage with empirical score functions is that pH, salt concentration and temperature can influence the measured binding constants significantly. This is often ignored when the datasets used for training the score functions are derived (Brooijmans and Kuntz, 2003). Examples of empirical score functions showing some promise include PLP (Gelhaar *et al.*, 1995; Gelhaar *et al.*, 1999), ChemScore (Eldridge *et al.*, 1997) and X-Score (Wang *et al.*, 2002). PLP uses a sum of pairwise linear potentials between ligand and protein heavy atoms with parameters dependent on interaction type. Each pair of interacting atoms is assigned one of three interaction types: donor and acceptor hydrogen bonding, repulsive donor-donor and acceptor-acceptor interactions and dispersion contacts. The ChemScore function includes hydrogen-bonding terms, terms accounting for coordinate bonding between the ligand and metal ions placed in the protein binding pocket, hydrophobic effects and the number of rotors, while the X-Score function contains a van der Waals interaction term, a hydrogen bonding term, a term representing the hydrophobic effect and a torsional entropy penalty.

Knowledge-based score functions are derived by statistical analysis of structural data alone, without reference to experimentally determined binding affinities. The frequency of occurrence of individual contacts is used as a measure of their energetic contribution to binding. The frequencies are compared to frequencies from a random or average distribution. A high frequency indicates an attractive interaction, while a low frequency indicates a repulsive interaction. The Potential of Mean Force (PMF) score function (Muegge and Martin, 1999; Muegge, 2000; Muegge, 2001) is an example of a knowledge-based score function. The PMF score function is a sum of distance-dependent interaction potentials for atom pairs, where both enthalpic and entropic effects are assumed to be included implicitly. In the DrugScore equation (Gohlke *et al.*, 2000), individual potentials for protein and ligand atoms that are dependent on the size of the solvent-accessible surfaces of the protein and the ligand that become buried upon complex formation are also included.

Recently, eleven score functions were compared using the same set of experimental structures (Wang *et al.*, 2003). X-Score and DrugScore were found to be the score functions most suited for

use with conformational sampling, since they produce a funnel-shaped energy surface for protein-ligand complexation. This leads to a relatively fast convergence to the global minimum. This study also indicated that a combination of several different score functions might be advantageous. X-Score, DrugScore and PLP were the score functions showing most promise, but only a moderate correlation was obtained between the predicted (using the experimentally determined conformation) and the experimental binding affinities for the 100 complexes included in the study. Most of the eleven tested score functions predict hydrophilic interactions better than hydrophobic interactions.

2.3.6 De novo ligand design

If a drug molecule with the desired properties cannot be found by virtual screening of structural databases, an alternative is to use *de novo* ligand design to construct drug candidate structures from a set of proposed functional groups. There are three main approaches to *de novo* ligand design: linking, growing and random connection (Figure 2.5). Molecular fragments placed at possible interaction sites in the receptor binding pocket can be used as starting points for all three approaches. Most of the random connection methods start from an initial "pool" of molecular fragments and construct ligands by making and breaking connections between the fragments. Random connection methods include methods using genetic algorithms. In the same way as in computational docking, score functions are used to guide the building of the ligand structures, and to estimate the binding affinity. Examples of *de novo* ligand design methods include DycoBlock (Liu *et al.*, 1999), which uses the linking approach, SPROUT (Gillet *et al.*, 1993; Gillet *et al.*, 1994), where the growing approach is used, LigBuilder (Wang *et al.*, 2000), where both growing and linking are possible, and ADAPT (Pegg *et al.*, 2001), which uses a genetic algorithm. Available *de novo* ligand design methods are listed elsewhere (Schneider and Böhm, 2002; Anderson, 2003).

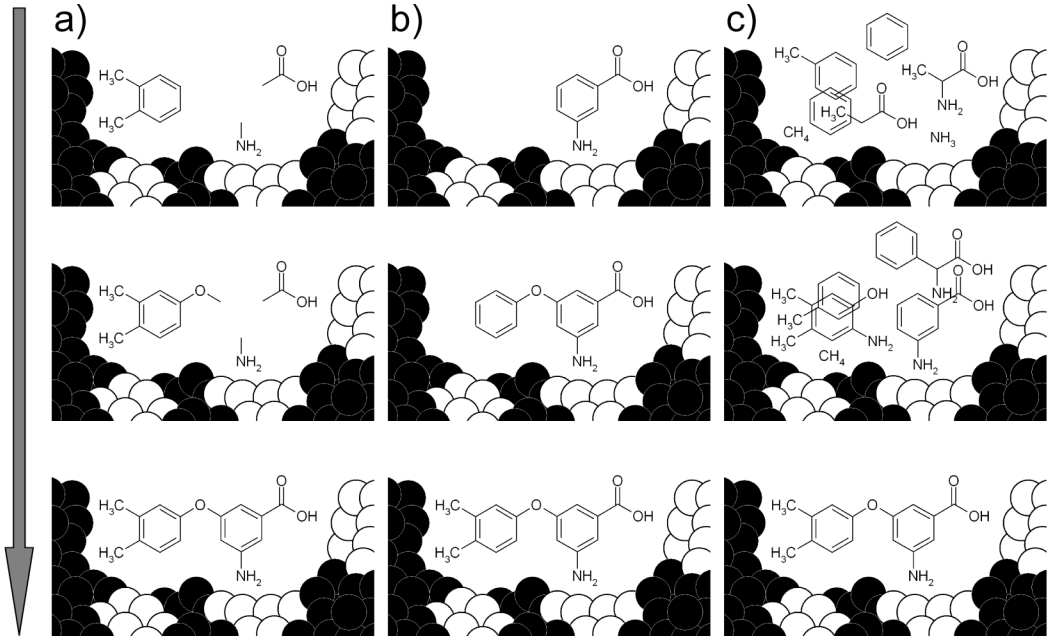


Figure 2.5. The three main categories of *de novo* ligand design methods (figure from Paper III). a) In the linking approach, molecular fragments placed close to important residues of the protein are connected to obtain a ligand. b) The growing approach starts from one fragment and connects fragments sequentially to it. c) Most of the random connection methods start from an initial "pool" of fragments and construct ligands by making and breaking connections between the fragments.

Few *de novo* ligand design methods exist that take factors such as synthetic accessibility, bioavailability and metabolic properties into account. Many ligand suggestions produced by these methods have large and complex structures. Recently, some programs have been developed that attempt to take such factors into account. An example is LigBuilder (Wang *et al.*, 2000), which uses a filter to make sure that the produced structures have reasonable ADMET (Absorption, distribution, metabolism, excretion and toxicity) properties. LigBuilder starts by analysing the protein binding pocket using three different probes, an ammonium cation (hydrogen donor), a carbonyl oxygen (hydrogen acceptor) and methane (hydrophobic group). The interaction energies between the probes and the protein are calculated using an empirical score function accounting for van der Waals interactions, hydrogen bonding, hydrophobic interactions and entropy loss due to freezing of rotatable bonds in the ligand. LigBuilder builds up ligands iteratively by using a library of organic fragments. Both growing and linking strategies are possible, and the construction process is controlled by a genetic algorithm.

Leads produced by a *de novo* ligand design method can also be evaluated for their likelihood of being orally bioavailable using the "Rule of five" (Lipinski *et al.*, 1997), which suggests upper limits for the number of hydrogen donors and acceptors, the molecular weight and the octanol/water partition coefficient. Rigidifying the lead can also produce a lower binding constant by decreasing the conformational entropy in the unbound state, and thereby decreasing the difference in entropy between the bound and the unbound state (Anderson, 2003).

In the same way as in molecular docking, protein flexibility is often ignored by *de novo* ligand design methods. However, a few methods exist that account for fluctuations in the protein structure. A recent version of DycoBlock, F-DycoBlock (Zhu *et al.*, 2001) uses multiple-copy stochastic molecular dynamics, while in the "Dynamic Pharmacophore Method" (Carlson *et al.*, 1999), pharmacophore models are determined for a large number of MD snapshots.

De novo ligand design has contributed to the development of several important drug leads (Sawyer, 2001). An important example is the discovery of STI-571, a selective inhibitor of Abelson (Abl) kinase, which is being used as a therapeutic agent against chronic myelogenous leukaemia (Schindler *et al.*, 2000; Capdeville *et al.*, 2002). Other examples include the development of antifungal agents (Ji *et al.*, 2003) and the design of aspartyl protease inhibitors. The aspartyl protease inhibitors were verified experimentally (Ripka *et al.*, 2001).

3 Molecular systems

The present work has focused on development of selective inhibitors for proteins that are involved in cancer development and metastasis. In this chapter, an introduction to the structure and function of the proteins that have been studied here will be given.

3.1 Protein kinases

The protein kinase superfamily is one of the largest protein families known to date. The protein kinases constitute a family of signalling proteins that is involved in a wide variety of biological functions. Protein kinases contribute to regulation and coordination of e.g. metabolism, gene expression, cell growth, cell motility, cell differentiation and cell division (Johnson *et al.*, 1996). The protein kinases are divided into two main groups: non-receptor protein kinases and receptor protein kinases.

Tyrosine protein kinases constitute a sub-family of the protein kinases, and are usually regulated by tyrosine phosphorylation (Hubbard and Till, 2000). Most protein kinases contain an activation segment that is about 25 residues long (Johnson *et al.*, 1996). This activation segment begins with a highly conserved DFG motif, and ends with a conserved APE motif. The region between the DFG and APE motifs is called the activation loop. With few exceptions, phosphorylation of tyrosine residues in the activation loop of tyrosine kinases leads to an increase in enzymatic activity (Hubbard and Till, 2000). Phosphorylation of tyrosines outside of the activation loop can negatively regulate kinase activity.

As illustrated in Figure 3.1, receptor tyrosine kinases (RTKs) consist of an extracellular portion that binds polypeptide ligands, a transmembrane helix, and a cytoplasmic portion that possesses tyrosine kinase catalytic activity (Hubbard and Till, 2000). The non-receptor protein kinases contain only a cytoplasmic part. The majority of RTKs are monomeric in the absence of ligand. Ligand binding to RTKs leads to receptor oligomerisation and tyrosine autophosphorylation. Autophosphorylation of tyrosine residues leads to increased kinase catalytic activity, and generation of docking sites for protein substrates. The RTKs catalyse the transfer of the γ phosphate of adenosine triphosphate (ATP) to the hydroxyl group of a tyrosine in a substrate protein. This triggers signalling cascades that participate in a large number of biological processes.

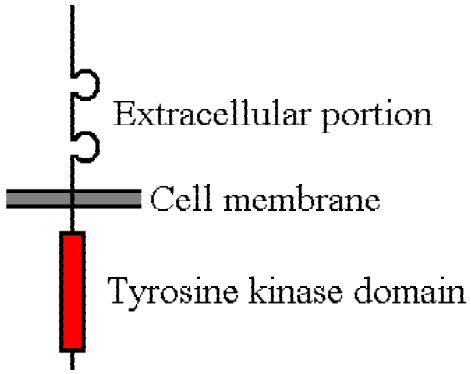


Figure 3.1. Illustration of the RTK structure.

Tyrosine phosphorylation in the activation loop of tyrosine kinases causes conformational changes in the activation loop (Pautsch *et al.*, 2001). Pautsch *et al.* describe the conformational changes of insulin-like growth factor 1 (IGF1) upon kinase activation. According to their work, the global conformational changes caused by kinase activation are triggered by the degree of

phosphorylation and are dependent on the conformation of the kinase activation loop. Crystal structures of the insulin receptor tyrosine kinase in both the active and the inactive form have also been determined (Hubbard *et al.*, 1994; Hubbard, 1997), and the conformational differences in the activation loop have been described. In the inactive form, the activation loop (residues G1149-L1170) occupies the active site, and thereby occludes substrate binding, while the triphosphorylated activation loop of the active form permits access to the binding sites for both ATP and protein substrates (Hubbard, 1997). The active conformation of the activation loop is very similar in all known structures of active kinases, but there is a large diversity in the conformations of this loop in inactive protein kinases (Schindler *et al.*, 2000). Several structural intermediates might exist between the inactive, dephosphorylated and the fully active, phosphorylated forms of protein kinases (Pautsch *et al.*, 2001).

3.1.1 Janus kinases

The Janus kinase (Jak) family consists of four known non-receptor tyrosine kinases (Tyk2, Jak1, Jak2 and Jak3) that play a critical role in initiating signalling cascades of a large number of cytokine receptors (van der Geer *et al.*, 1994; Ihle *et al.*, 1995; Pellegrini and Dusanter-Fort, 1997; Richter *et al.*, 1998). All Jak family kinases possess a carboxyl-terminal tyrosine kinase catalytic domain, a central pseudo-kinase domain, and a large amino-terminal region, which has been subdivided into five Jak homology regions (JH7 to JH3) based on sequence conservation (Harpur *et al.*, 1992; Richter *et al.*, 1998). The structure of the Jaks is illustrated in Figure 3.2. The pseudo-kinase domain is enzymatically non-functional, but may participate in regulation of kinase activity (Pellegrini and Dusanter-Fort, 1997; Hubbard and Till, 2000). Recently, homology modelling has been utilised to predict the structural mechanism by which this pseudo-kinase domain contributes to regulation of the Janus kinases (Lindauer *et al.*, 2001). In contrast to most other cytoplasmic protein tyrosine kinases, the Janus kinases have no Src homology (SH) domains (Ihle *et al.*, 1995).

Figure 3.2. Illustration of the structure of the Janus kinases. The JH domains are shown in grey, the pseudo-kinase domain is shown in blue, and the kinase catalytic domain is shown in red.

Ligand binding to cytokine receptors activates the Janus kinases through the specific and non-covalent association of these kinases to the intracellular region of the receptors (Pellegrini and Dusanter-Fort, 1997). The Jak activation is mediated by ligand-induced receptor oligomerisation (Schlessinger and Ullrich, 1992; Heldin, 1995; Hubbard and Till, 2000). The JH domains have been shown to be the parts of the Jaks that are associated with the cytoplasmic domains of cytokine receptors (Pellegrini and Dusanter-Fort, 1997; Richter *et al.*, 1998; Yan *et al.*, 1998). The Janus kinases are activated by e.g. the type I interferons (IFNα/β and γ), the interleukins (IL2-7, IL-10 and IL-12), growth hormone (GH), erythropoietin (Epo), granulocyte-specific colony-stimulating factor (G-CSF), granulocyte-macrophage colony-stimulating factor (GM-CSF), leukaemia inhibitory factor (LIF), ciliary neurotrophic factor (CNTF) and prolactin (Ihle *et al.*, 1995; Carter-Su and Smit, 1998; Richter *et al.*, 1998).

Activated Janus kinases autophosphorylate (Pellegrini and Dusanter-Fort, 1997), and phosphorylate the cytokine receptors with which they are associated. This provides binding sites for the Signal Transducers and Activators of Transcription (STAT) transcription factors (Hubbard and Till, 2000). The Jaks catalyse phosphorylation of the STAT proteins (Xuan *et al.*, 2001). Seven STAT isoforms, STAT1-4, STAT5A-B and STAT6, are known. Following phosphorylation on tyrosine residues, the STATs form homo- or heterodimers (Heldin, 1995), which are translocated

into the cell nucleus. The STAT proteins then bind to DNA (deoxyribonucleic acid), and activate gene transcription (Ihle *et al.*, 1995). The Jak-STAT signalling cascade has been shown to contribute to growth and survival of e.g. human multiple myeloma cells (Anderson, 1999), acute lymphoblastic leukaemia (Meydan *et al.*, 1996) and a variety of other malignancies (Wang *et al.*, 1999; Lindauer *et al.*, 2001). This makes the Janus kinases potential targets for new cancer therapies. Inhibiting binding of ATP to the kinases is one way to interrupt these signalling cascades. ATP analogues are generally non-selective, but the development of inhibitors like STI-571 (Schindler *et al.*, 2000) shows that ATP binding sites can be used as targets for selective drugs. At the present time none of the Janus kinases have experimentally determined 3D structures (Berman *et al.*, 2000). It is therefore interesting to predict the 3D structures of these tyrosine kinases, and especially the ATP binding regions.

3.1.2 Fibroblast growth factor receptor

Fibroblast growth factor receptor (FGFR) is a receptor protein kinase that is involved in e.g. angiogenesis, the process by which new capillaries are formed from pre-existing vessels. Angiogenesis is involved in embryo development, ovulation and wound repair. The normal regulation of angiogenesis is governed by a fine balance between factors that induce the formation of blood vessels and those that halt or inhibit the process (Gray *et al.*, 1998). Numerous factors that regulate angiogenesis have been identified, including members of the fibroblast growth factor (FGF) family, vascular endothelial growth factor (VEGF), angiogenin, transforming growth factor (TGF) α and β, platelet-derived growth factor (PDGF), platelet-derived endothelial cell growth factor (PDECGF), tumour necrosis factor (TNF) α, interleukins, chemokines and angiopoietins (Folkman and D'Amore, 1996; Gray *et al.*, 1998).

Angiogenesis is also essential for growth and metastasis of tumours (Hanahan and Folkman, 1996; Kumar and Fidler, 1998). In the same way as normal cells, cancer cells are dependent on blood supply for survival. Pathological angiogenesis (abnormal rapid proliferation of blood vessels) is also involved in a large number of other diseases, such as diabetic retinopathy, atherosclerosis, rheumatoid arthritis, age-related macular degeneration and psoriasis (Klagsbrun and Edelman, 1989; Klagsbrun and D'Amore, 1991; Pepper, 1996; Kuiper *et al.*, 1998; Szekanecz *et al.*, 1998; Tolentino and Adamis, 1998). This makes the angiogenic factors and their receptors potential targets for development of new therapeutic agents.

3.2 Lectins

The lectins constitute a class of specific carbohydrate-binding proteins distinct from both sugar-specific enzymes and antibodies, which contributes to regulating many biological processes (Gabius, 1988). The natural ligands of lectins include the sugar moieties of cell glycoproteins, glycolipids and proteoglycans. The specific protein-sugar interaction between lectins and carbohydrates constitutes a reciprocal recognition system, utilised e.g. for information storage and signal passage by the immune system. Lectins are previously known to be overexpressed by mammalian malignant cells compared to normal ones (Gabius, 1988; Raz *et al.*, 1990; Vodovozova *et al.*, 2000). This overexpression can e.g. be used for targeting of anticancer drugs coupled to macromolecular carriers to tumours with help of specific carbohydrate ligands (Vodovozova *et al.*, 2000; David *et al.*, 2004).

Selectins contain lectin domains essential for carbohydrate binding (Revelle *et al.*, 1996; Stahn *et al.*, 1998). Sialyl Lewis x ($SiaLe^x$) is a terminal unit of cell-surface glycoproteins and glycolipids, and has been identified as a common ligand for E-, P- and L-selectin (Huwe *et al.*, 1999). Binding of $SiaLe^x$ to E- or P-selectin mediates the early stage of the inflammatory response, leading to the

rolling of neutrophil cells on the endothelium, followed by the recruitment of these cells to inflamed tissue (Huwe *et al.*, 1999). Inhibition of neutrophil adhesion to endothelium is an attractive approach to controlling inflammation-mediated diseases such as rheumatoid arthritis and psoriasis (Robinson and Stephens, 1992). The selectins are also known to mediate the adhesion of cancer cells to endothelium (Klopocki *et al.*, 1998). This suggests that inhibition of the binding of $SiaLe^x$ to selectins might be utilised in cancer therapy, as well.

4 Summary of the papers

First, a brief overview of the presented work will be given, and the different papers will be placed into context. How the different topics are linked together is also shown. The different parts of the work are dealt with in more detail in section 5.

The first step in a drug design process is generation of a structural model of the target system. Homology modelling of protein structures has achieved increasing attention, since the amount of sequence data increases much faster than the amount of available structural data (Berman *et al.*, 2000; Marti-Renom *et al.*, 2000). The development of new methods that can utilise the structural information that is present in the homology models without being misled by the structural errors can significantly increase the effectiveness of a rational drug design process, as well as the number of possible drug targets that can be considered. The use of homology modelled protein structures in drug design is reviewed in Paper III. The accuracy of the homology models is an important issue, determining to a large extent the reliability of the results from the drug design process. Even the most robust drug design methods are dependent on a certain level of accuracy of the structural models used. Hence, it is important to assure that in each case, the best homology model possible is being used. It is therefore interesting to examine the quality of homology models, and to be able to predict the quality of future homology models. In Paper I, the accuracy of homology models has been evaluated, and a method for prediction of model quality based on the sequence alignment between the target and template has been developed. A large number of homology models of protein kinase structures were generated, and the accuracy of the homology models was evaluated by comparison to available X-ray structures of the targets. Based on this homology model quality data, a method for prediction of homology model accuracy with multivariate regression was developed. This method predicts the model accuracy directly from the amino acid sequence alignment of the target sequence to the template used for the homology modelling. Hence, no information about the 3D structure is needed to predict the model quality for new homology models. This method can be used to assure that the optimal templates and alignments are chosen, so that the best possible homology model is generated. It is also useful for identification of regions of the protein structure that are difficult to model, as well as errors in the alignment. This method has been applied to the protein kinase family, but can easily be extended to other protein families.

Given a reliable model of the 3D structure of a protein, the binding site properties have to be analysed, in order to find drug candidates with matching properties. As described in Chapter 2.3.3, numerous methods exist for analysis of protein binding sites and detection of possible interaction sites for drug candidates. However, few drug design methods exist that are robust against the additional structural error that is introduced as a result of using homology modelled protein structures instead of the more accurate experimental structures that have been used traditionally. This work has focused on the development of drug design methods utilising gaussian functions to represent atomic properties. This gives a smoothing of the molecular descriptions, and the idea is that this smoothing will give a more robust representation of the molecular properties and interactions than methods that utilise more detailed descriptions of the protein binding sites (McGann *et al.*, 2003). Detailed information requires accurate structural models, otherwise it will be misleading. Our goal was to describe the most important factors involved in protein-ligand interaction, using as few variables as possible. In Paper II, a new method for mapping of protein binding site properties, called Protein Alpha Shape Similarity Analysis (PASSA) is introduced. Once a map of the protein binding site is generated, this information can be used to search databases of already existing drugs to find structures that match the protein binding site properties, and to generate new structures having the desired properties using *de novo* ligand design. With PASSA, structural regions that can be utilised to inhibit proteins selectively can be identified. PASSA is especially useful when combined with for example MCSS, which can be used to suggest small molecular fragments that can bind at the identified selective interaction sites in the protein binding

pocket. PASSA is an effective method for comparison of the binding sites of several related proteins, since the same grid can be used to map the binding sites of all proteins, and multivariate data analysis tools can be used to analyse the results. This allows for effective identification of the interaction sites that can be utilised to achieve selectivity. Using knowledge about already identified selective ligands, binding site descriptions generated with PASSA can also be used to model selectivity within a protein family, and to identify the protein-ligand interactions that are important for selectivity. This information can then be used to design new selective ligands for other proteins in the family. The PASSA method has been used to suggest functional groups for a selective inhibitor of Tyrosine kinase 2 (Tyk2) (Paper II), and to model selectivity within the protein kinase family (Paper VII). The results presented in Paper VII demonstrate that the PASSA method may be used to predict the activities of a number of ligands towards a set of closely related protein targets. This makes PASSA a promising method in screening for side effects. The suggested functional groups from Paper II have been used further in Paper VI, where the NCI database was searched for existing drugs having Tyk2 inhibitory activity. The proposed functional groups were also utilised to construct inhibitor candidate molecules by *de novo* ligand design.

When a set of possible drug candidates is found, their affinity to the target receptor has to be estimated, in order to identify the most active compounds. In this work, two different approaches to estimation of the binding energy between a receptor and a ligand were used. In Paper IV, the binding free energy was calculated from "first principles", using the sum of the electrostatic contribution to the solvation energy, the Coulombic energy and a term representing the hydrophobic effect, estimated by calculation of solvent-accessible surface areas. This PB/SA approach is too time-consuming to be effective in large-scale virtual screening, but is a useful tool for optimisation of the functional groups of a known drug lead. In Paper IV, the interactions between the receptor kinase fibroblast growth factor receptor 1 (FGFR1) and a known inhibitor were studied by computational sensitivity analysis, and several improvements of the inhibitor were suggested. The results show that computational sensitivity analysis is an effective method for gaining insight into which ligand groups that have the largest contributions to binding, and which groups that should be modified in order to increase binding affinity. A comparative database analysis of almost 400 protein kinases was also carried out in order to identify groups on the inhibitor that should be modified to increase selectivity. In Paper V, a new docking method is introduced, that utilises gaussian property distributions to describe the receptor and ligand properties. This docking method uses an empirical score function to estimate the binding affinity. The score function was trained on 218 X-ray structures of protein-ligand complexes for which the binding affinity is known, and evaluates the match between the lipophilicity and hydrophilicity of the receptor and the ligand, in addition to describing van der Waals effects. For the ligand, information about hydrogen donors and acceptors was also included. By using gaussian functions representing hydrophilicity and lipophilicity, in addition to describing van der Waals effects, we aim to describe protein-ligand interactions better than methods that only account for steric clashes. The use of gaussian property distributions also makes this docking method suitable for use with homology models. This docking method has been shown to work well for ligands that bind in a well-defined cavity in the protein, but may fail for ligands forming interactions with the outer protein surface. Hydrogen atom positions and partial charges are not taken into account in the calculations, so the binding affinity might be underestimated for ligands forming many hydrogen bonds or ion bonds with receptor atoms. Solvent effects are also ignored. The accuracy of our docking method is lower than that of more time consuming methods that use fewer approximations. However, this docking method has been shown to be fast, and is therefore well suited for virtual screening, where the main purpose is to rank a large number of ligands according to binding, and identify a small set of promising drug candidates that is worth working with further. In Paper V, the performance of our gaussian-based docking method was compared to that of MOE-Dock, using Tabu search for conformational searching in both cases. MOE-Dock performed better than our method in reproducing ligand X-ray structures, but our method used only 10% of the computational time compared to MOE-Dock.

MOE-Dock succeeded in identifying the correct conformation for a larger number of ligands than our method, but except for the absolutely correct predictions, the results were comparable for the two methods.

The interactions between a set of carbohydrate ligands and E-selectin were studied with computational docking. The results obtained with our docking method were compared to those obtained using the much more time consuming simulated annealing approach in MOE-Dock. Both methods failed in reproducing experimental binding affinities for the carbohydrates. This is not surprising, since the carbohydrate ligands are very flexible. This leads to a high probability of getting trapped in local minima. The carbohydrate ligands also bind to the outer surface of E-selectin, through a large number of hydrogen bonds. This makes our docking method unsuitable in this case. The high flexibility of the carbohydrates and the large number of possible carbohydrate receptors might make them unsuitable as selective drugs. Other ligands, such as peptides, might therefore be more interesting. A set of peptide ligands was also docked in E-selectin using our docking method. The results clearly demonstrate that our method only works in cases where the ligand binds in a cavity in the protein structure.

The gaussian-based docking method developed here was applied in the design of selective inhibitors of Tyk2 in Paper VI. Here this docking method was used to dock both structures resulting from screening of the NCI database, and drug candidate molecules generated by *de novo* ligand design. The selectivity of the compounds was tested by computational docking in seven other protein kinase structures. The results from the docking of the compounds from the NCI database were compared to the results obtained with MOE-Dock. The two docking methods ranked the structures differently, but produced the same conclusion, namely that none of the compounds in the NCI database can inhibit Tyk2 selectively. One compound was found to inhibit Tyk2 and insulin receptor tyrosine kinase selectively, and five of the drug candidates from the *de novo* ligand design seem promising as selective Tyk2 inhibitors.

The different parts of the present work and how they are linked together are illustrated in Figure 4.1.

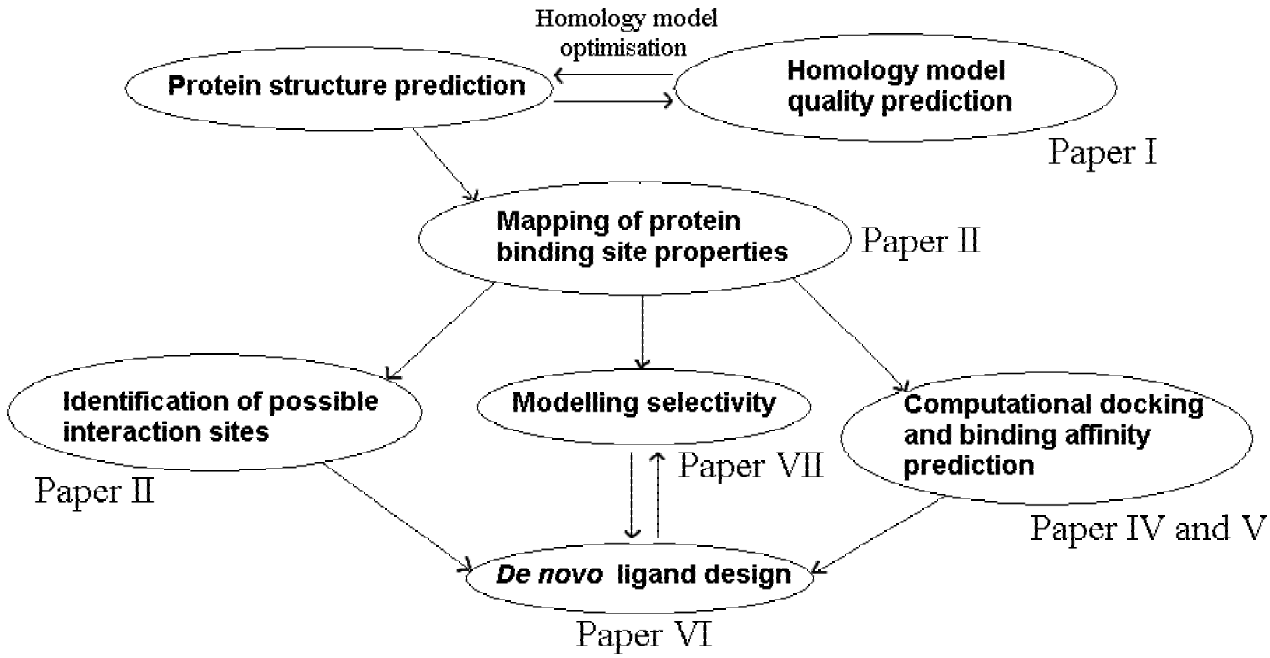


Figure 4.1. Overview of the presented work.

5 Case studies

5.1 Verification and prediction of homology model accuracy

A new method for prediction of homology model quality with multivariate regression has been developed (Paper I). This method predicts the model accuracy directly from the amino acid sequence alignment of the target sequence to the template used for the homology modelling. Hence, no 3D structure information is needed for prediction of the accuracy of new homology models, once the regression model has been trained. Here, this method has been applied to the protein kinase family, but it can easily be extended to other protein families. Prediction of the model quality that can be expected, given a specific relationship between the primary structure of the target and template proteins is useful for evaluating whether a homology model can be generated that can provide useful information. Depending on the specific project, different levels of accuracy are needed. A large variety of methods and programs exist for predicting homology model quality, most of which operate on the 3D structure models (Chapter 2.3.2). Few methods exist that predict the model quality prior to the actual model building. Hence, the method presented here provides a new way to predict homology model accuracy. Using this method, time can be saved compared to using methods that predict the model accuracy after the homology model has been generated, since generation of low-quality models can be avoided. One can also find out in advance, whether it is possible to generate a model with the required accuracy. With this method, insight into possibilities for improving the model quality by modelling different domains separately and correcting alignment errors can also be gained. This information can also be obtained using other methods, such as methods based on 3D profiles, but with these methods, the homology model has to be generated before any information about the expected accuracy can be obtained. A method for prediction of local backbone structural deviation in homology models has been developed, that can be used to evaluate the quality of a preliminary homology model (Cardozo *et al.*, 2000). This method is useful for optimising a homology model by providing information about for example loop boundaries. While our method provides a quick estimate for the expected overall homology model accuracy based on the sequence alignment, this method provides a more detailed description of the probability that a given backbone segment is predicted accurately. Our method is most useful in the initial stages of the modelling, to evaluate e.g. whether it is possible to generate a useful homology model for a given application.

The results from the regression analysis (see Paper I and Appendix 3-4) show that the residuals from prediction of the model quality for new homology models can be used to identify proteins that are difficult to model with homology modelling due to large deviations from the other members of the protein family, as well as provide useful information about alignment errors. The regression coefficients from the analysis can also be used to identify problem regions in the protein structure and alignment errors. Hence, the method presented here can be used to assure that the optimal templates and alignments are chosen, so that the best possible homology model is generated. Using the method developed here, the model quality can be predicted based on a sequence alignment, possible problem regions can be identified using for example the residuals from the regression analysis, and corrections to the sequence alignment can be made. The impact of the alignment corrections on the expected homology model accuracy can then be tested by predicting the accuracy for the new alignment. Hence, this method is an effective tool for optimising the sequence alignment prior to generating the homology model. Different homology modelling methods may also perform differently for a given sequence alignment. Using the approach developed in this work, regression models can be made that can predict the homology model quality resulting from several different homology modelling methods. This can guide the choice of modelling method, and assure that the best possible homology model is generated, given a certain sequence alignment. This method can also be used to determine in what cases several templates should be utilised for the

homology modelling, and in what cases a single template provides sufficient information. In Paper I, possibilities for improving the model quality by combination of several homology models are discussed.

A set of 292 homology models of protein kinases for which experimental structures are available in the PDB was generated with WHAT IF (Vriend, 1990), using a modelling pipeline for automatic homology modelling (Liu *et al.*, in preparation). The target-template sequence identities ranged from 14 to 80%. The resulting homology models were verified by comparison to the experimental structures of the targets. RMSD values (separate overall Cα, Cβ and heavy atom RMSDs) between the model structures and experimental structures of the target proteins, and differences in inter-residue contact areas between the models and the target X-ray structures (unnormalised CAD numbers) were used as measures of the model quality. To describe the sequence alignment between the target and template proteins, a matrix of alignment score profiles was generated. Each element in this alignment score matrix contains the value of the Point Accepted Mutation (PAM) 250 similarity matrix (Schwartz and Dayhoff, 1978) for a pair of amino acids that correspond to each other in the sequence alignment (Figure 5.1).

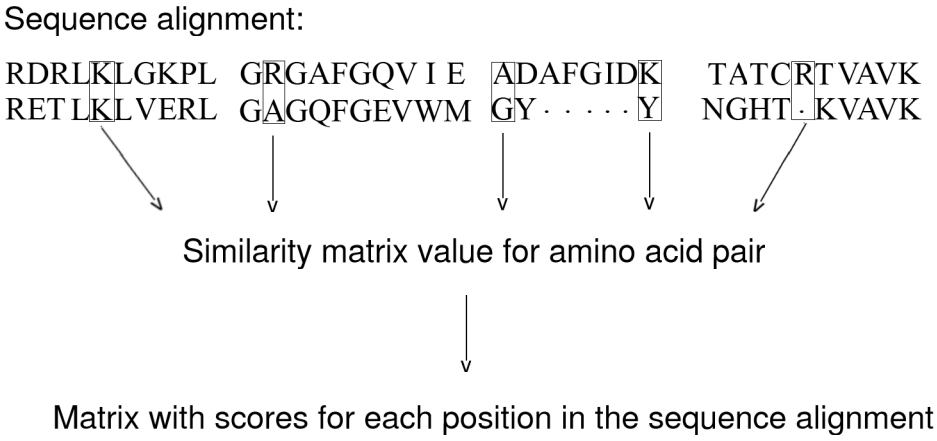


Figure 5.1. Generation of alignment score profiles.

The model quality dataset was analysed with PLS regression using alignment score profiles, sequence identity and number of non-modelled residues as independent variables, as illustrated in Figure 5.2. The results for the contact area error are shown in Figure 5.3.

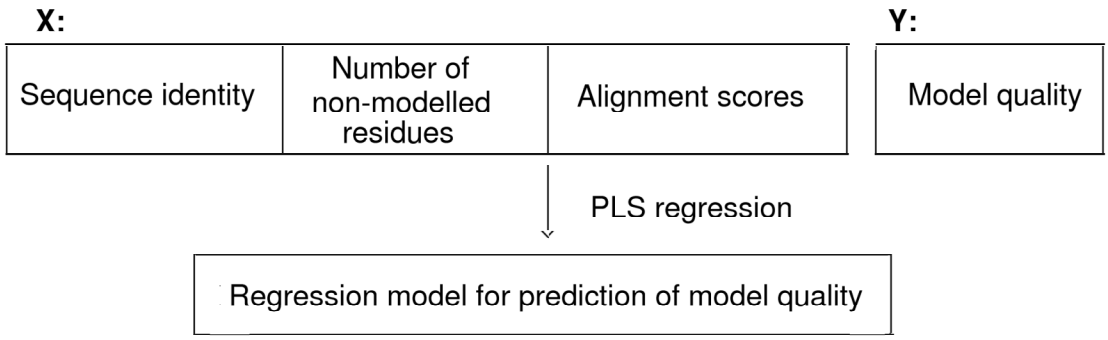


Figure 5.2. Multivariate analysis of the model quality data using alignment scores, sequence identity and number of non-modelled residues as independent variables.

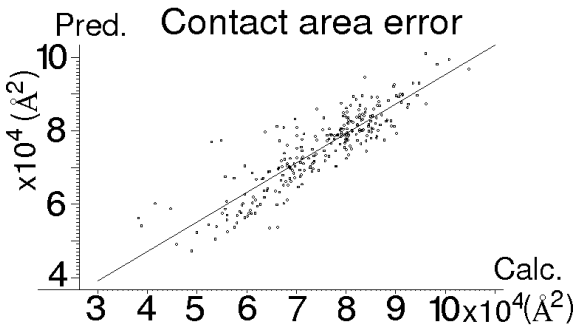


Figure 5.3. Predicted (from cross-validation) versus calculated contact area error ($Å^2$) for the 292 homology models.

Figure 5.3 shows a good correlation between the predicted and the calculated contact area error for the 292 homology models. The correlation coefficient (q) is 0.88. Hence, the alignment score profiles generated from the PAM250 similarity matrix provide sufficient information about the similarity between the targets and templates to predict the homology model accuracy.

5.2 Mapping protein binding site properties

A new method for protein binding site analysis based on gaussian functions, called Protein Alpha Shape Similarity Analysis (PASSA), has been developed in Paper II. The use of gaussian functions to describe the atomic properties makes PASSA especially suited for use with homology models, since this gives a smoothing of the molecular surface representation. PASSA focuses on design of protein inhibitors that are selective. PASSA starts by identifying the positions of geometrical objects known as alpha spheres. An alpha sphere is a sphere that contacts four atoms on its surface and has no atoms in its interior. Alpha spheres are determined geometrically, using only the positions and radii of the protein heavy atoms, and are classified as hydrophobic or hydrophilic depending on the protein atoms they contact. Large alpha spheres are present on the outer protein surface, while the very small alpha spheres are placed in small crevices in the protein structure. The middle-sized alpha spheres are most interesting in drug design, since they are typically placed in cavities large enough to hold a drug molecule. Hence, only the middle-sized alpha spheres are used in PASSA. Centres of middle-sized alpha spheres have been found to correspond well with the placement of atoms in bound ligands (Liang *et al.*, 1998).

In PASSA, gaussian functions are centred at dummy atoms placed at each alpha sphere centre, and at all protein atoms. A 3D grid is placed around the binding site of the protein, and in each grid point the sum of the contributions from all gaussian functions is computed. The contributions from gaussian functions centred at alpha spheres classified as hydrophobic define the hydrophobicity field, while the hydrophilicity field is defined by gaussian functions centred at hydrophilic alpha spheres. Ligand inaccessible volume is defined by the gaussian functions centred at protein atoms.

The use of gaussian functions with a very simple partitioning according to the hydrophilic or hydrophobic nature of the alpha spheres reduces some of the problems associated with force field models, like GRID (Goodford, 1985). The potential energy functions used in most force field models have steep derivatives close to atomic nuclei and singularities in the atomic nuclei, since the van der Waals interactions are estimated using the Lennard-Jones potential, E_{LJ} (Equation 5.1), and the calculation of the electrostatics is based on Coulombs law (Goodford, 1985; Leach, 2001).

$$E_{LJ} = \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} \tag{5.1}$$

r_{ij} represents the interparticle distance, while A and B are constants depending on the depth of the potential well and the interparticle distance for which the energy is zero. Steep derivatives and singularities are not associated with gaussian functions. In molecular dynamics, soft-core potentials have been utilised to overcome these problems (Huber *et al.*, 1997; Tappura *et al.*, 2000; Hornak and Simmerling, 2003).

Analysis of data from gaussian fields typically produces contour plots that are less fragmented and easier to interpret than those produced using force field models (Böhm *et al.*, 1999). PASSA is also more computationally efficient than most force field methods and methods based on solving the Poisson-Boltzmann equation. Rule-based methods for mapping protein binding sites have the disadvantage that their performance is dependent on the similarity of the target protein to the proteins included in the dataset used to derive the statistical rules. This is not the case for PASSA.

In the same way as with GRID, the data produced with PASSA can be compared for a large number of related proteins since a regular grid is used to compute the data. This is a great advantage when the purpose is to design inhibitors that are selective. In PASSA, the structural regions that can contribute to selectivity are identified directly, using discriminant partial least squares regression (DPLSR). In DPLSR, the dependent variables are indicator variables, and the regression seeks to describe the separation of the samples into classes defined by these indicator variables, using the data contained in the X-matrix. In PASSA, the X-matrix contains the gaussian property field data. The protein structures included in the analysis are divided into classes, one containing the target protein, and one or more classes containing related proteins for which a low affinity is desired. When protein structures built by homology modelling are used, it is advantageous to use more than one model of each protein, particularly if several templates of comparable sequence identity are available. In this case, several alternative homology models of comparable accuracy are possible, and including more than one model might give a better representation of the properties of the protein. The regression coefficients from the DPLSR indicate structural regions where each class of proteins has properties that separate these proteins from the other proteins included in the analysis. Interactions between the protein and inhibitor groups placed in areas of high regression coefficients can contribute to selectivity. The regression coefficients can be visualised as contours in the original 3D space of the protein structures. The results from PASSA can be combined with for example MCSS, to identify functional groups that can contribute to selective inhibition of the target protein. The PASSA method is illustrated in Figure 5.4.

Other methods also exist that utilise sphere-based approaches to analyse protein structures. Examples of such methods include PASS (Brady and Stouten, 2000), APROPOS (Peters *et al.*, 1996) and CAST (Liang *et al.*, 1998). However, these methods focus on localisation of the protein binding site, not on identification and characterisation of interaction sites for selective ligands, like PASSA. These methods use distributions of spheres to detect cavities in the protein structure. They do not use gaussian functions and summation over a 3D grid to obtain a continuous description of the binding sites, and provide no direct way to compare the properties of several proteins to single out the interaction sites that can contribute to selective binding of an inhibitor. A knowledge-based method for mapping protein binding sites has been developed, that utilises gaussian functions to generate smooth propensity maps from scatter data stored in the ISOSTAR database (Nissink *et al.*, 2000). The propensity maps reflect the probability of finding an interacting group close to a given central group. The main disadvantage of this method compared to PASSA is the dependency on structural data for parameterisation. An advantage with this method is that propensity maps for a large number of different molecular fragments can be generated. PASSA generates only two different property maps, one for hydrophilicity and one for hydrophobicity.

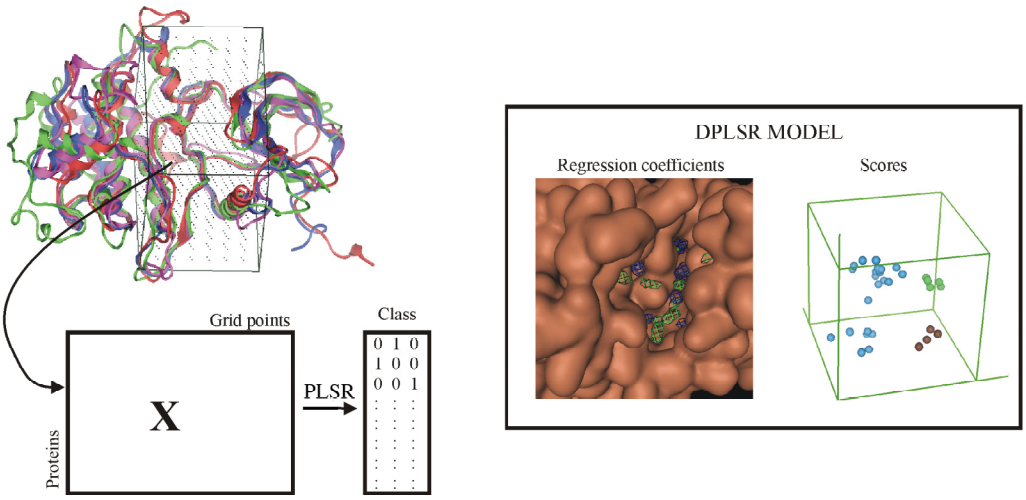


Figure 5.4. Illustration of the PASSA method. The values of the gaussian property fields are used as independent variables in a DPLSR using a protein class matrix as dependent variables. DPLSR represents differences between the target protein and the other proteins in the study as a vector of regression coefficients that can be visualised in the 3D space of the protein structures. The distribution of the protein structures in the score plot shows the separation of the proteins according to class memberships.

The performance of PASSA was verified by testing whether residues known to interact with selective inhibitors are among the residues identified by PASSA to have properties that are unique to the target protein. STI-571 is a selective inhibitor of Abl kinase (Zimmermann *et al.*, 1997; Schindler *et al.*, 2000). In Figure 5.5, the regression coefficients for Abl kinase are visualised together with the X-ray structure of STI-571 in complex with Abl kinase (PDB entry 1IEP).

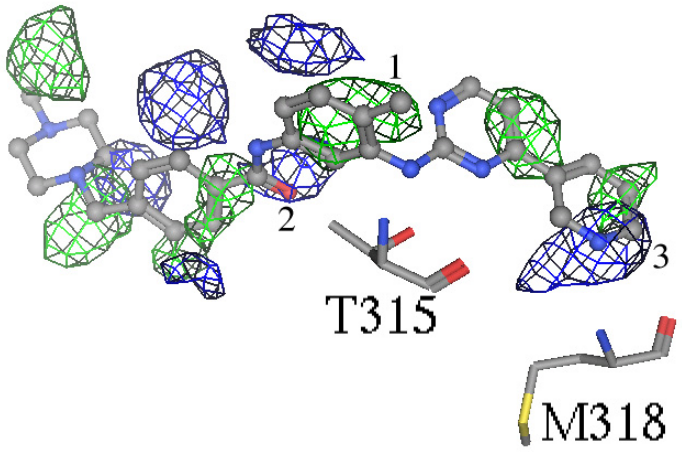


Figure 5.5. Plot of the regression coefficients for the hydrophilicity (blue) and the hydrophobicity (green) from the DPLSR mapped back onto the grid surrounding the ATP binding pocket of Abl kinase in complex with STI-571 (PDB entry 1IEP). Residues of Abl kinase known to be important for selectivity are shown. STI-571 is rendered in "ball and stick", while the protein residues are rendered in "stick". The phenyl-moiety of STI-571 interacting with T315 in Abl kinase is numbered "1", the carbonyl group of STI-571 is numbered "2", while the nitrogen atom interacting with M318 is indicated by the number "3".

Figure 5.5 shows that the areas of high regression coefficients for the hydrophobicity correspond well with the hydrophobic parts of STI-571. The phenyl-moiety of STI-571 known to interact with T315 in Abl kinase (Schindler *et al.*, 2000) is placed in an area where Abl kinase is particularly hydrophobic compared to the other proteins. The interaction between T315 and STI-571 is known to be important for selectivity (Schindler *et al.*, 2000; Gorre *et al.*, 2001). According to our results, Abl kinase is particularly hydrophilic in the region close to the carbonyl group of STI-571, and around the nitrogen interacting with M318. A similar analysis was carried out using a homology model of Jak2 in complex with a docked conformation of AG490. AG490 inhibits Jak2, but none of the other proteins included in this analysis (Meydan *et al.*, 1996; Bright *et al.*, 1999; Kirken *et al.*, 1999; Wang *et al.*, 1999; Xuan *et al.*, 2001). The regions where Jak2 is particularly hydrophobic according to our PASSA results correspond well with the hydrophobic parts of AG490, and likewise for the hydrophilicity. The fact that the interactions between Abl kinase and STI-571, and between Jak2 and AG490 were identified by PASSA indicates that this approach is well suited for identification of interaction sites that can contribute to selectivity. This makes PASSA a useful method in the design of selective drugs.

Figure 5.6 shows the results from the PASSA for Tyk2. According to our results, Tyk2 has three unique hydrophobic pockets that can be utilised to achieve selectivity towards Tyk2 (shown in Figure 5.6 A, B and C, respectively). Similar analysis for the hydrophilicity identified useful hydrogen acceptors and donors close to these pockets. Here, the results from PASSA have been combined with MCSS.

According to our results, interactions with hydrogen acceptors or donors on Y955, E1053, D1062 and S1063 can be utilised to achieve selectivity towards Tyk2. Fragments from MCSS placed in regions of high regression coefficients were used to indicate possible functional groups for a selective Tyk2 inhibitor. These results can be used as a starting point for combinatorial library generation, database searching and *de novo* ligand design.

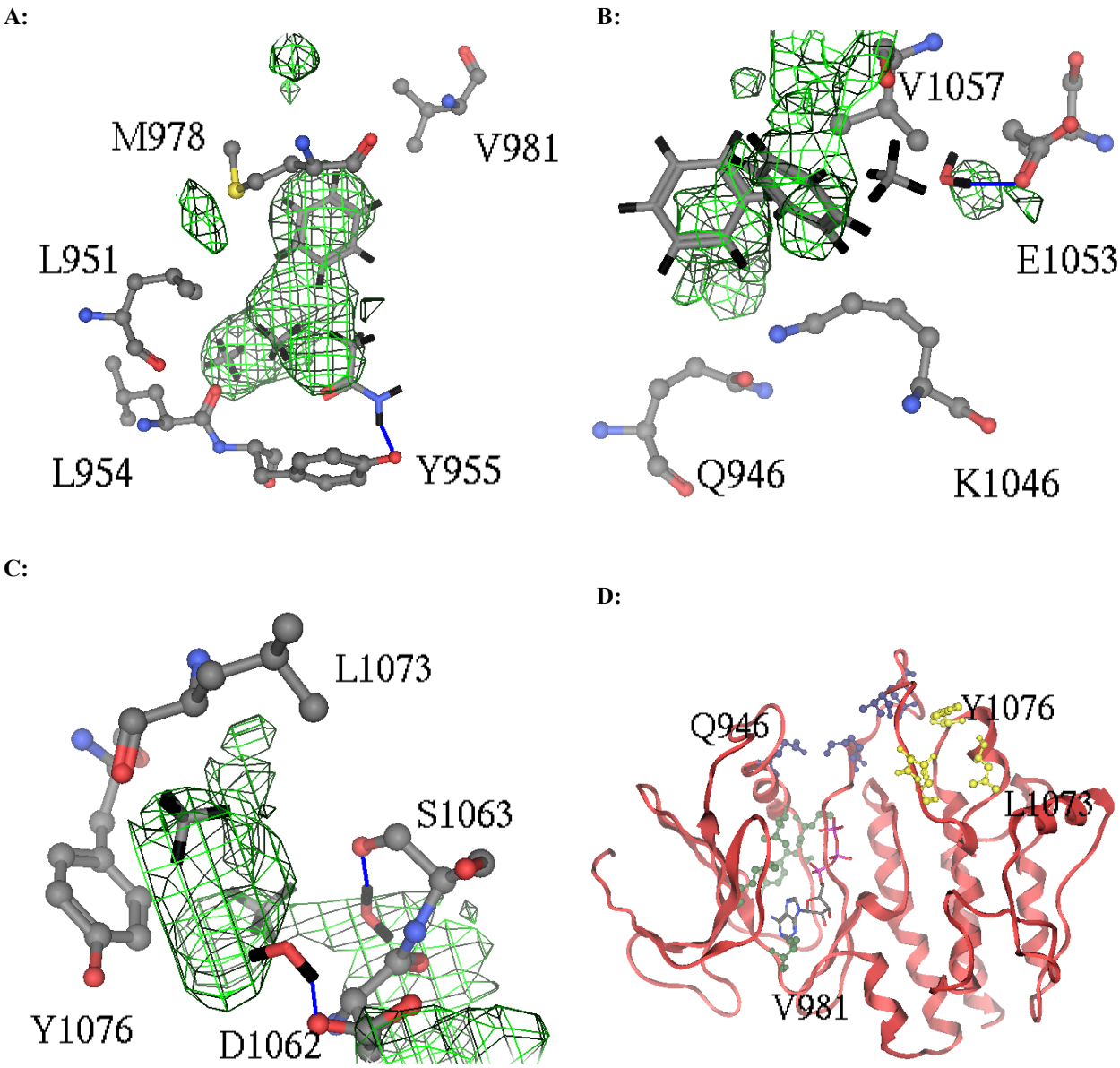


Figure 5.6. **A-C:** Plots of the regression coefficients for the hydrophobicity (green) from the DPLSR mapped back onto the grid surrounding the model of the ATP binding pocket of Tyk2. Residues identified by PASSA to be unique to Tyk2 are shown, together with selected fragments from MCSS. Possible hydrogen bonds between MCSS fragments and hydrogen acceptors and donors on Tyk2 identified to be unique are shown as blue lines. **D:** The residues from Fig. 5.6 A (green), 5.6 B (blue) and 5.6 C (yellow) shown together with the result from computational docking of ATP in the Tyk2 model.

5.3 Modelling interactions between proteins and drug candidates

5.3.1 Computational analysis of the interactions between the angiogenesis inhibitor PD173074 and fibroblast growth factor receptor 1

The effects of the potent angiogenic factors FGF and VEGF are mediated through the cell surface receptors fibroblast growth factor receptor and vascular endothelial growth factor receptor (VEGFR), that possess intrinsic protein tyrosine kinase activity (Mohammadi *et al.*, 1998). A compound of the pyrido[2,3-d]pyrimidine class (PD173074) has been reported, that selectively inhibits the tyrosine kinase activity of FGFR and VEGFR (Mohammadi *et al.*, 1998). This inhibitor contains a dimethoxyphenyl group that occupies a pocket in the ATP-binding cleft that is not utilised by ATP. Mohammadi *et al.* suggest that this group is important for the selective binding of this inhibitor. In Paper IV, the interactions between the angiogenesis inhibitor PD173074 and FGFR1 were studied using computational sensitivity analysis, and functional groups of the inhibitor that are important for binding affinity and selectivity were identified. Several improvements of the inhibitor were also suggested.

The basic idea of computational sensitivity analysis is similar to mutational analysis of recombinant proteins, in which one examines whether a particular feature of an amino acid affects a protein property by mutating the amino acid into another one that no longer contains the feature. In a computational sensitivity analysis, one "mutates" parameters of a molecular model, such as atomic partial charges and dipole moments of functional groups, to examine the significance of these features in affecting binding affinity (Wong *et al.*, 1998). In Paper IV, information from computational sensitivity analysis was utilised to identify the parts of a drug lead being most important for binding and the parts that should be modified to increase the binding affinity. This approach has been applied earlier to study the binding of balanol and a peptide inhibitor to protein kinase A (Wong *et al.*, 2001; Gould and Wong, 2002), and has been shown to be an effective tool for optimising a drug lead. Information from computational sensitivity analysis can also guide the design of focused chemical libraries that may produce more useful new hits, and the parts of a drug lead that have been identified to be useful for binding can guide the construction of pharmacophore models for mining new drug leads from small-molecule libraries.

The significance of a model parameter in affecting binding energy was analysed by calculating derivatives of the form $d\Delta G/d\lambda_i$ estimated by

$$d\Delta G/d\lambda_i = (\Delta G_{binding}(\text{mutant})-\Delta G_{binding}(\text{wildtype}))/\Delta\lambda_i \quad (5.2)$$

where $\Delta G_{binding}$(wildtype) represents the binding free energy between PD173074 and FGFR1, and $\Delta G_{binding}$(mutant) is the corresponding quantity when the collective charge or dipole moment of an atom or functional group is changed by $\Delta\lambda_i$. This is referred to as a "mutation". The collective charge equals the charge of an atom if only one atom is involved or the sum of the charges within a group if a group of atoms is involved. To analyse the effect of a "mutation", we calculated $(d\Delta G/d\lambda_i)\lambda_i$, where λ_i represents the collective charge or dipole moment of an atom or a group of atoms. A negative $(d\Delta G/d\lambda_i)\lambda_i$ indicates that the "mutation" leads to an improvement of the binding affinity. We also estimated the effects of pairwise interactions on binding by calculating second derivatives of the form $(d^2\Delta G/d\lambda_i d\lambda_j)\lambda_i\lambda_j$. These second derivatives estimate pairwise interactions as in double mutagenesis experiments.

Computational sensitivity analysis utilises much more accurate predictions of the binding affinities between the protein and the ligands than what is utilised in most docking methods. In this work, a PB/SA approach was applied. No conformational search was used, the functional groups of the X-ray structure of the FGFR1 inhibitor were "mutated", and the ligand conformation was kept fixed. Energy minimisation was used only in cases where new groups were introduced. This

approach is useful for fine tuning the activity of a known drug lead, but it is inefficient in a docking study where a large number of different ligands are evaluated, and the correct binding modes are not known in advance. In this case, more approximate methods for estimating the binding affinity have to be applied. One advantage of using this method is that the binding affinity is calculated from "first principles", that is, without reference to experimental data. The performance of empirical and knowledge-based score functions is dependent on the size and diversity of the set of structures used to derive the equations. This is not the case for first-principles methods. The predictive ability of our computational model was tested by comparing the calculated results to experimental binding affinities to FGFR1 for eight ligands, including the angiogenesis inhibitor PD173074. The correlation between the estimated binding affinities and the experimental IC_{50} values was quite good, with a correlation coefficient of 0.8. This indicates that this method can provide useful information about groups of the inhibitor PD173074 that should be kept and parts that should be modified to improve binding affinity.

When designing drugs targeting the ATP binding pocket of protein kinases, specificity is especially important to consider, since the ATP binding pocket is a common feature of all protein kinases. A comparative database analysis of almost 400 protein kinases was carried out to gain insight into how PD173074 may be modified to improve selectivity. The results from this study and the computational sensitivity analysis are summarised in Figure 5.7.

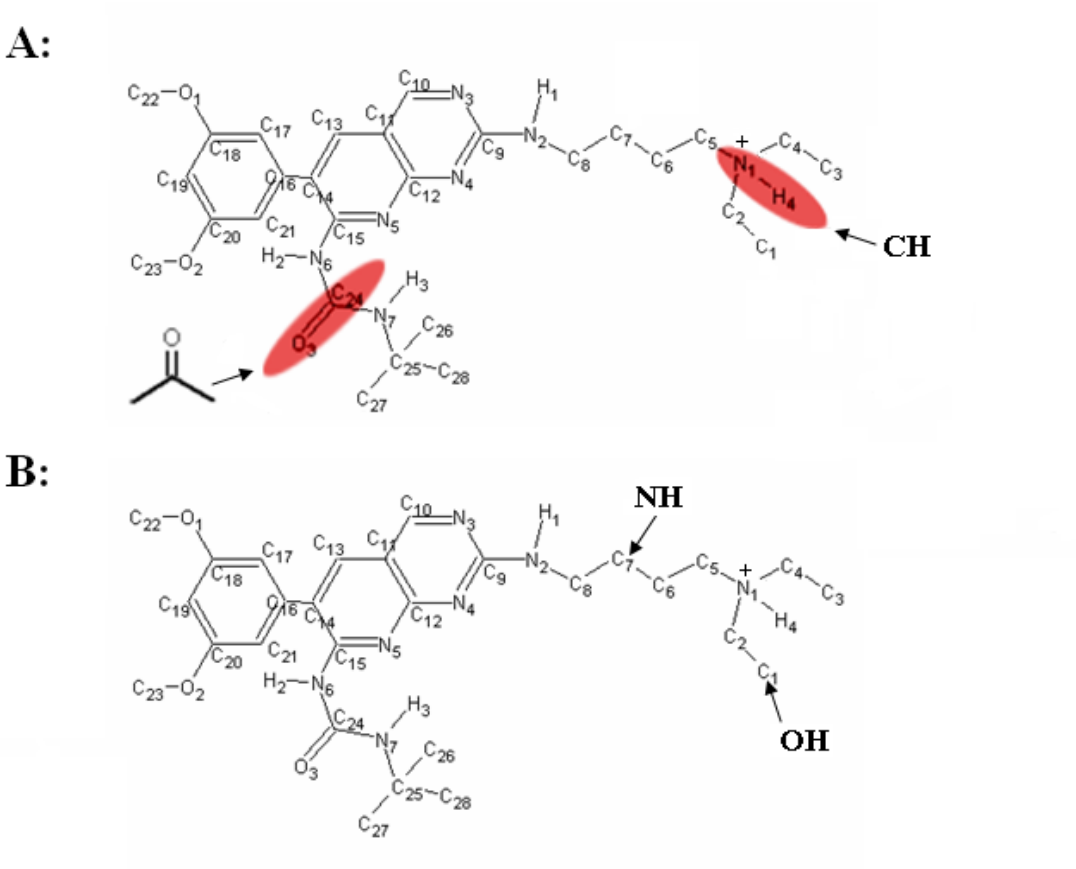


Figure 5.7. Summary of the results from the computational sensitivity analysis and the comparative database analysis. **A:** Changes that may increase binding affinity. The functional groups that should be modified are indicated in red, while the suggested modifications are shown close to these groups. A hydrophobic group should replace the positively charged diethylammonium group, while the carbonyl group should be replaced by e.g. $CH(COCH_3)$ to push out the carbonyl oxygen. **B:** Changes that may increase selectivity. The CH_2-group at C_7 should be replaced by an NH-group, and an OH-group should replace the CH_3-group at C_1.

The positively charged diethylammonium group was found to diminish binding. Unless it is important to use this group to improve aqueous solubility, it may be better to replace this positively charged ammonium group with a hydrophobic group. The polarity of the NH-group closest to the pyrido[2,3-d]pyrimidine (at N_6) seems useful, but the carbonyl group should be replaced with e.g. a $CH(COCH_3)$ group to improve binding affinity. In addition, the CH_2-group at C_7 (Figure 5.7) should be replaced by an NH-group, and an OH-group should replace the CH_3-group at C_1 to improve selectivity. Our analysis also indicated that the dimethoxyphenyl ring should be modified in order to improve binding affinity. It seems favourable to keep the oxygens of the methoxyl groups but the methyl groups may be replaced with other non-polar groups. According to the database analysis, selectivity may also be achieved by modifying this part of the molecule. The results from the database analysis suggested that introducing functional groups ortho to the pyrido[2,3-d]pyrimidine ring might improve selectivity.

5.3.2 A new gaussian-based docking method suitable for use with homology modelled proteins

A new empirical score function for estimation of binding affinities to a receptor using gaussian property distributions for both the protein and the ligands has been developed (Paper V). The score function evaluates the match between the lipophilicity and hydrophilicity of the receptor and the ligand, in addition to describing van der Waals effects. For the ligand, information about hydrogen acceptors and donors was also included. The protein binding site was described using gaussian property fields calculated in the same way as in PASSA (Paper II), while the ligand properties were described using gaussian property fields similar to those used in the 3D-QSAR method CoMSIA (Klebe *et al.*, 1994). The match between the protein and ligand properties was evaluated by calculating the products between the protein and ligand fields in each grid point. The product values were summed over all grid points, to give the parameters used in the score function. The score function was trained on 218 X-ray structures of protein-ligand complexes for which experimental binding affinities are available, using PLS regression. While most existing docking methods use the same score function both in the conformational search and for binding affinity prediction, we use a faster version of the score function in the conformational search to limit the computational time (Paper V).

Hydrogen atom positions and partial charges are not described by our score function. This is of course a significant limitation of the method, but we wanted a fast score function that could be used for virtual screening, and that is robust against the structural errors present in homology models. Including information about e.g. hydrogen atoms and partial charges requires accurate protein structures. Hence, we had to balance between robustness and accuracy. Our goal was to capture the main features of binding, and omit variables that are sensitive to errors in the structural models.

This method only works for ligands that bind in a well-defined binding pocket, since the properties of the protein are described using alpha spheres that are placed in cavities in the protein structure. However, when the purpose is to design highly specific ligands, deep pockets in the protein structure are more interesting than interactions on the outer protein surface (Zavodszky *et al.*, 2002). A sufficiently large interaction surface is needed to achieve high affinity, and specificity is more easily obtained within a cavity, which already imposes geometric constraints (Sotriffer and Klebe, 2002). Hence, it is easier to construct a ligand that binds selectively in a small, well-defined pocket. A large variety of compounds can attach to the outer surface of a protein, due to the large number of possible interaction sites. Hence, to achieve selectivity, one has to construct ligands large enough to interact with several different interaction sites simultaneously. The size and flexibility of these compounds complicate the use of computational docking to predict their activities, and the suitability of these compounds as drugs might be low due to factors such as bioavailability.

Since we include no information about hydrogen atom positions, we are unable to represent direction-specific hydrophilic interactions. Hydrophobic interactions are not direction-specific.

Hence, our score function predicts hydrophobic interactions better than hydrophilic interactions. Many other score functions for computational docking predict hydrophilic interactions better than hydrophobic interactions (Wang *et al.*, 2003). Hence, it would be interesting to combine our score function with other score functions to improve the predictive ability.

The fact that our score function is based on gaussian property distributions makes it relatively robust against small structural errors. As mentioned earlier, gaussian functions give a smoother representation than e.g. force field models. Gaussian functions have neither steep derivatives nor singularities (Paper III). Since this score function is robust against small structural variations, including protein flexibility is less important than in many other docking methods.

As described in Chapter 2.3.4, other gaussian-based docking methods also exist, that are suited for use with homology models. The method reported by McGann *et al.* (2003) only accounts for shape, and is trained on a much smaller set of protein-ligand complexes than our docking method. By including hydrophilicity and hydrophobicity in addition to van der Waals effects, and by using a larger training set, we aim for an improved description of protein-ligand interactions. Another gaussian-based method that also accounts for hydrophilicity and hydrophobicity in addition to shape has been reported (Schafferhans and Klebe, 2001). As for our method, gaussian functions are used to represent the physicochemical properties of the receptor and the ligand, and the overlap between the functional descriptions of the receptor binding site and the ligand is optimised. This method uses interaction sites identified by the *de novo* ligand design program LUDI (Böhm, 1992 a, b) to generate a description of the protein binding site. As the method developed by McGann *et al.* (2003), this docking method is also trained on a relatively small set of complexes. Hence, this method is sensitive to deviations between the target system and the training set. It is also indicated that this method predicts both hydrophobic interactions and electrostatics insufficiently (Schafferhans and Klebe, 2001). Their results indicate that gaussian functions are too soft to model electrostatics sufficiently. In contrast to our docking method, this method is dependent on assignment of charges and protonation states. This information is not trivial to generate, especially not for protein models with potential inaccuracies, such as homology models. One advantage with this method is that it can take several different homology models into account simultaneously by an averaging of the property densities of the models. This increases the robustness of this method.

Because of the very simplified description of the protein-ligand interactions, the accuracy of our new score function can not be compared to that of score functions that take e.g. electrostatics into account. However, the speed of our calculations makes this method an effective tool for pre-screening for virtual drug design. In virtual screening, the purpose is to identify a set of promising drug candidates from a large collection of ligand structures. Hence, the binding affinity has to be estimated for a large number of structures. This makes computational efficiency an important factor to consider. In virtual screening, the purpose is not necessarily to predict the absolutely correct binding modes for all ligands, or predict the binding affinity with a high level of accuracy. It is most important to effectively separate the active compounds from the non-active ones. The correct ranking of the promising drug candidates and the correct binding conformations can be found with more accurate and time consuming methods, once the number of structures to consider has been reduced.

Figure 5.8 shows an example of a gaussian property description of a known protein-ligand complex. Only the hydrophilicity field for the protein is shown, together with the ligand structure. As seen from the figure, the hydrophilic groups of the ligand and the hydrophilicity field of the protein match to a high degree for this complex. The hydrophilic phosphate groups are for example surrounded by the contours of the hydrophilicity field of the protein. This indicates that gaussian property fields generated by PASSA provide useful descriptions of the protein binding site properties that can be used in computational docking, where the purpose is to maximise the complementarity between the protein and the ligand in a conformational search.

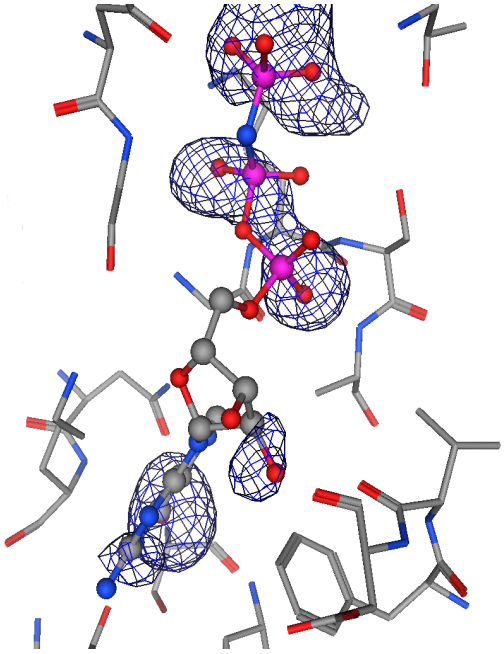


Figure 5.8. The gaussian hydrophilicity field for the protein in PDB entry 1AGP shown together with the ligand X-ray structure. The hydrophilicity field is indicated by the blue mesh. The ligand is rendered in "ball and stick", while the protein binding site residues are rendered in "stick".

The performance of our new empirical score function was verified in a docking analysis using all protein-ligand complexes in the training set. The ability of our method to reproduce the experimental structures and binding affinities was tested. The results were compared to the results obtained with MOE-Dock (MOE, 2002). Tabu search was used for the conformational search with both methods. The results are given in Figure 5.9. The predicted binding affinities for the docked conformations resulting from docking with our gaussian-based score function are plotted against the experimental binding affinities in Figure 5.10.

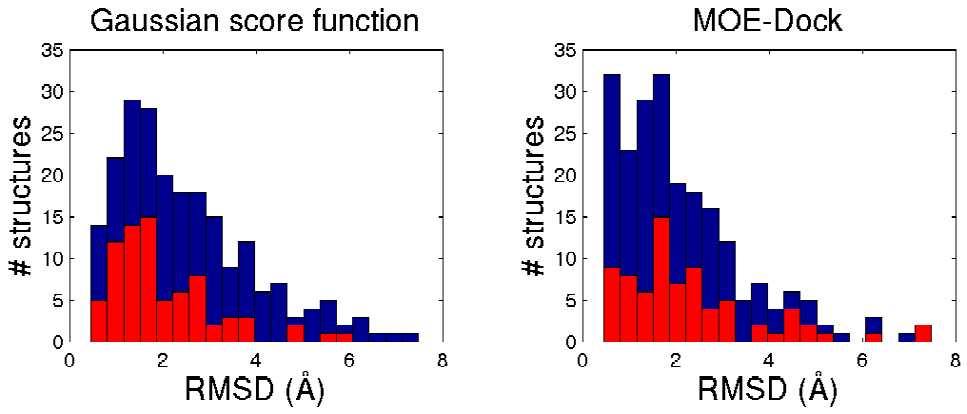


Figure 5.9. **Left:** Histogram over RMSD values between the X-ray ligand structures in the training set and the ligand structures resulting from 10 Tabu runs of 1000 iterations each using our gaussian-based score function. The fraction of the complexes having experimental binding affinities below –40 kJ/mol is shown in red. This docking procedure used ~5 minutes per molecule, on average.
Right: Histogram over RMSD values between the X-ray ligand structures in the training set and the ligand structures resulting from 10 Tabu runs of 1000 iterations each with MOE-Dock. The fraction of the complexes having experimental binding affinities below –40 kJ/mol is shown in red. This docking procedure used ~50 minutes per molecule, on average.

The histograms in Figure 5.9 show that MOE-Dock performs better than our docking method. Using our score function we were able to get 102 of the 218 ligand conformations within 2 Å RMSD of the X-ray structure, while docking with MOE-Dock resulted in 120 of the ligand conformations within 2 Å RMSD. However, the two histograms showing the distribution of obtained RMSD values for the two docking methods are almost identical, except for the first column, representing the number of structures within 0.6 Å RMSD of the X-ray structure. Hence, the main difference in accuracy between the two methods is in the prediction of the absolutely correct conformations. In virtual screening, the main goal is to identify possible drug candidates. Hence, a reasonable prediction of the bound conformation might be sufficient. The histogram to the left in Figure 5.9 indicates that our docking method performs best for ligands having high affinity for the receptor (shown in red). For MOE-Dock, this relationship is not that clear. This indicates that our score function is well suited for virtual screening, where the purpose is to separate high affinity compounds from non-active ones, and to find a reasonable ligand conformation for a large number of protein-ligand complexes.

Our method is much more computationally efficient than MOE-Dock. Our method used approximately 5 minutes per molecule, while MOE-Dock used 50 minutes per molecule, on average. The level of accuracy of our score function might not be sufficient to give reliable results alone, but since our docking method is very fast, it is well suited for pre-screening and generation of starting conformations for more accurate docking. This docking method might also be complementary to other docking methods, since our results show that our method predicts hydrophobic interactions better than hydrophilic interactions. The opposite is true for many other docking methods (Wang *et al.*, 2003). Since combination of several score functions for computational docking has been shown to improve the results (Wang *et al.*, 2003), a combination of our score function with other score functions that predict for example hydrogen bond formation better might be advantageous. The fact that MOE-Dock and our method did not succeed in reproducing the X-ray structures for the same ligands indicates that the two methods are complementary to a certain degree.

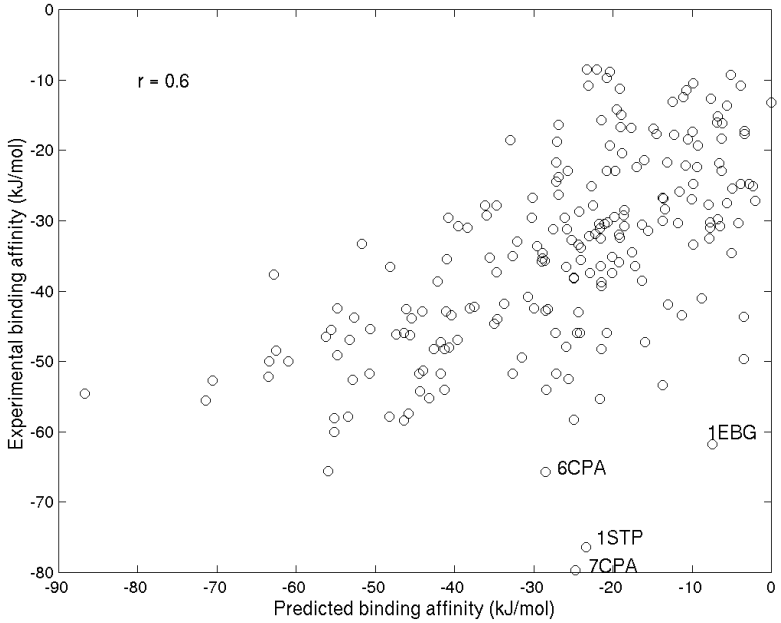


Figure 5.10. Predicted binding affinities for the ligand structures resulting from the docking analysis with our gaussian-based score function plotted against the experimental binding affinities. The docking analysis of the X-ray structures with PDB entries 1STP, 6CPA, 7CPA and 1EBG resulted in false negatives.

The correlation between the experimental and the predicted binding affinities obtained using our gaussian-based docking method is only 0.6. This is probably caused by the low number of variables included in the score function, and the simplified molecular representation used. In our score function, neither partial charges, nor hydrogen atom positions are taken into account. Solvent effects are also ignored. We wanted a robust score function including only the variables that are most important for binding. A higher correlation could probably have been obtained by including more variables in the score function, but this would make the method less robust against structural errors. Including information with large errors might be worse than leaving the variables out. The false negatives shown in the lower, right hand part of Figure 5.10 result from underestimation of the strength of ion bonds (1EBG) and hydrogen bond formation (1STP) between the ligand and the receptor. Since information about partial charges and hydrogen atoms is not included in our score function, the binding affinity is underestimated in cases where the ligand makes ion bonds to the receptor and when hydrogen bond formation is very important for ligand binding. Hydrogen bond formation is dependent on the directions in which hydrogen atoms point, and even though we include information about hydrophilicity and hydrogen donors and acceptors, we are not able to fully represent hydrogen bond formation using our simplified molecular description. The ligands in PDB entries 6CPA and 7CPA contain groups that protrude towards the solvent. Our description of the protein binding site is based on the positions of alpha spheres. Alpha spheres can only represent ligand atoms bound in a protein cavity. Hence, interactions on the outer surface of the protein are ignored, and the contributions from the protruding groups to the binding affinity are not included. This, combined with the exclusion of solvent effects probably causes the underestimation of the binding affinities of 6CPA and 7CPA.

Our score function is trained on a large and diverse set of protein-ligand complexes. Most existing score functions have been trained on smaller and more homogenous sets of structures. Our training set includes both small, rigid ligands, and relatively large and flexible ligands such as peptides. Predicting the binding modes of peptides has been a great challenge in computational docking. However, our method was for example able to reproduce the experimental conformations of three different peptides containing seven, eight and nine amino acids, respectively (PDB entries 8HVP, 1HHK and 1HHH). The RMSD values between the docked and the experimental conformations of these three peptides were 0.94 Å, 0.99 Å and 1.5 Å, respectively. This indicates that our method will be a useful supplement to existing docking methods, and can make important contributions to structure-based drug design.

5.3.3 Computational docking of carbohydrate ligands and peptides in E-selectin

Computational docking methods use many approximations to the system under consideration, and clearly have their limitations. The three computational docking examples presented here (unpublished results) did not give satisfying results, and illustrate some of the limitations of these methods. In these examples flexible carbohydrates and peptides binding to the outer surface of the protein were used. The presented results illustrate that methods based on calculation of alpha sphere positions are only suitable for ligands that bind in cavities in the protein structure. Carbohydrates bind to their receptors through a large number of hydrogen bonds. Hence, docking methods that predict hydrogen bond formation insufficiently (such as our gaussian-based docking method) are not suitable for use with these compounds. Compounds like peptides, with many rotatable bonds, are also known to represent a challenge in computational docking. The computational details and experimental binding affinity data are given in Appendix 1 and 2, respectively, and the results from the docking calculations are also given in Appendix 2.

A set of 34 fluorescent carbohydrate probes has been tested for binding to mammary adenocarcinoma cells (Vodovozova *et al.*, 2000). This analysis revealed the tetrasaccharide $SiaLe^x$ as the ligand with the highest affinity to the cancer cells. We contacted Vodovozova and co-

workers, and obtained the original dataset from their lab. In this dataset a discontinuous scale from zero to five was used to rank the carbohydrates according to their affinity to the mammary adenocarcinoma cells. The affinity of the carbohydrates to the cancer cells was estimated by microscopic evaluation of fluorescence emitted from the bound carbohydrate ligands. We have docked this set of carbohydrate ligands in E-selectin, a receptor that is overexpressed on various cancer cell lines (Gabius, 1988). Two different docking methods were used: docking by simulated annealing in MOE (Hart and Read, 1992), and our gaussian-based docking method presented in Paper V. The results from the docking calculations were compared to the experimental affinities of the ligands to the cancer cells.

The estimated docking energies (the sum of the electrostatic and the van der Waals interaction energy between the ligand and the target and the intramolecular energy of the ligand) from the simulated annealing docking of the 34 carbohydrates in E-selectin are plotted against the observed binding affinities to the mammary adenocarcinoma cells in Figure 5.11.

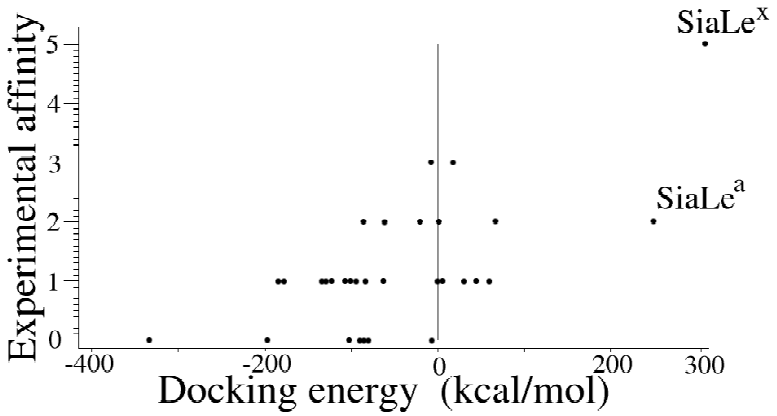


Figure 5.11. Docking energy from docking with simulated annealing in MOE plotted against the experimentally determined binding affinity. The correlation between the estimated and the observed affinities is 0.66.

The results in Figure 5.11 show that there is a significant correlation between the estimated and the experimental affinities, but a high experimental affinity should be associated with a low docking energy. The results here show the opposite. The main reason is that the values of the ligand intramolecular energies are very high for some of the compounds, especially for $SiaLe^x$ and Sialyl Lewis a ($SiaLe^a$). For most of the tri- and tetrasaccharides, the ligand intramolecular energies have values larger than 200 kcal/mol. These energies are lower for the ligands containing fewer ring structures. The values of the electrostatic and van der Waals interaction energies are comparable for all ligands studied here. The electrostatic energies typically have values around –200 kcal/mol, while a typical value for the van der Waals energy is 10 kcal/mol. This leads to positive docking energies for some compounds. This is probably caused by the high flexibility of these compounds. This docking method does not seem to be suitable for this kind of compounds, since it only separates the large, flexible ligands from the smaller ones. For carbohydrates containing up to four saccharide rings, a large number of local energy minima exist. As mentioned earlier, getting trapped in local minima with high-energy transition-state barriers is a common problem with simulated annealing. This might have caused the high intramolecular energies for these ligands. Recently, docking methods based on stochastic tunnelling have been developed to overcome some of these problems (Wenzel and Hamacher, 1999; Todorov *et al.*, 2003).

Since only interactions with E-selectin were considered in the docking experiments, some of the deviations between the estimated and the observed affinities might be due to binding of the carbohydrates to other receptors, such as P- and L-selectin. It is difficult to model binding to a complete cell system by modelling only one receptor-ligand interaction. During the simulated annealing calculations, solvent effects and receptor flexibility were not taken into account. Hence,

this model uses many approximations to the real system. The experimental affinities are also just a sensoric ranking of the compounds regarding to microscopic evaluation of fluorescence.

The fact that we were not able to reproduce the ranking between the carbohydrate ligands in this docking study was one of the reasons why we decided to develop a new docking method, based on gaussian property distributions like those used in PASSA (Paper II) and CoMSIA (Klebe *et al.*, 1994). Our hypothesis was that gaussian-based methods would work better on flexible structures, because of their robustness against small structural errors. The docking calculations with the gaussian-based docking method were performed in the same way as described in Paper VI, and the results are shown in Figure 5.12. Lately, the X-ray structures of E-selectin and P-selectin in complex with $SiaLe^x$ have been published (Somers *et al.*, 2000). These structures were not available when these docking calculations were done. A comparison of our docking results to the structure in PDB entry 1G1T showed that we were not able to find the correct binding mode for $SiaLe^x$ with either MOE-Dock or our gaussian-based docking method.

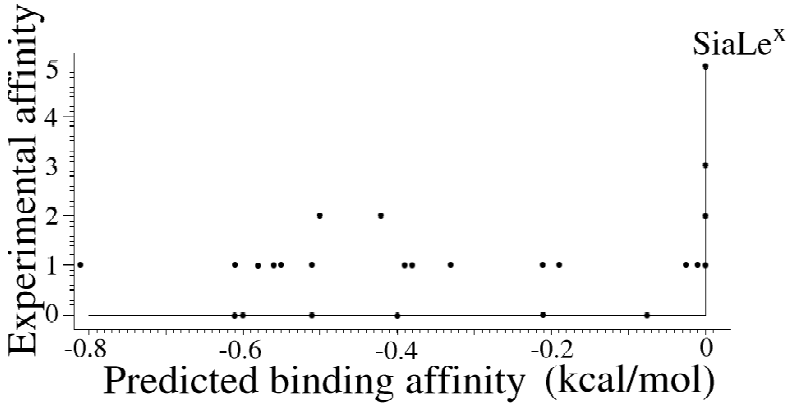


Figure 5.12. Estimated binding affinities from docking with the gaussian-based docking method plotted against the experimentally determined binding affinities. The correlation between the estimated and the observed affinities is 0.45.

The results in Figure 5.12 show that we were not able to reproduce the ranking of the carbohydrate ligands using our gaussian-based docking method. A large fraction of the ligands (for example the high-affinity ligand $SiaLe^x$) is predicted to have a binding affinity of zero to E-selectin, which indicates that they are placed outside the binding site on E-selectin, that is, outside the grid used to estimate the binding affinities. This illustrates that our docking method is not suitable for use with these compounds. One reason might be that hydrogen bond formation is important for binding of carbohydrates to their receptors. As mentioned earlier, our docking method does not include information about hydrogen atom positions. In addition to this, these carbohydrate ligands bind on the surface of the protein, not in a well-defined binding pocket (Graves *et al.*, 1994; Ng and Weis, 1997; Somers *et al.*, 2000; PDB entry 1G1T). The X-ray structure of E-selectin in complex with $SiaLe^x$ is shown in Figure 5.13. As discussed earlier, our docking method uses alpha spheres placed in cavities in the protein structure to represent the protein binding site, and interactions on the surface of the proteins are not taken into account. Solvent effects are also ignored.

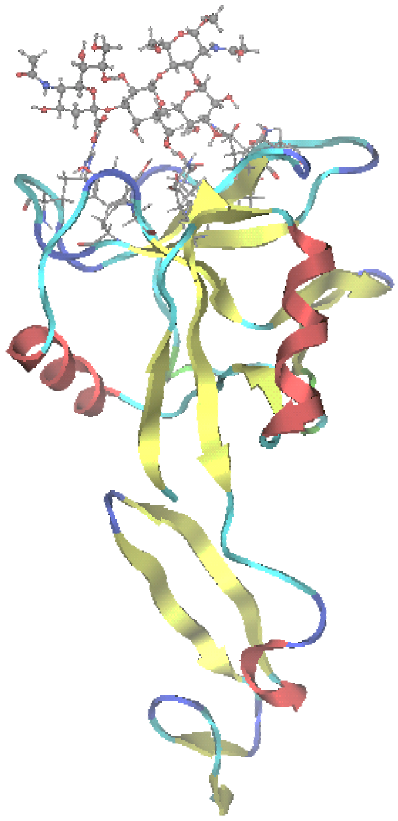


Figure 5.13. The X-ray structure of E-selectin in complex with SiaLe^x (PDB entry 1G1T). The ligand is rendered in "ball and stick", while the residues of E-selectin within 3 Å distance of SiaLe^x are rendered in "stick".

The high flexibility of the carbohydrates and the large number of possible carbohydrate receptors might cause selectivity problems. Other ligands, such as peptides, might therefore be better suited as selective drugs. A set of 25 peptides consisting of from ten to eighteen amino acids has been tested for binding to E-selectin (Martens *et al.*, 1995). In the same way as the carbohydrate ligands described above, these 25 peptides were docked in the X-ray structure of E-selectin, using our gaussian-based docking method. The predicted binding affinities are plotted against the experimental IC_{50} values (Martens *et al.*, 1995) in Figure 5.14.

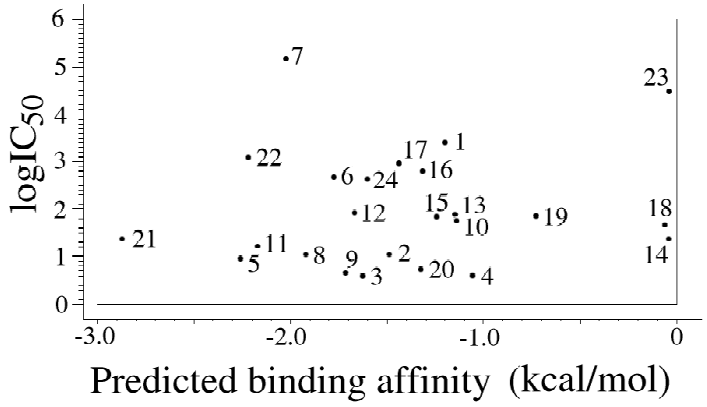


Figure 5.14. Predicted binding affinities for the peptides vs. logaritmised IC_{50} values. Peptide number 25 was kept out of the plot due to the very high predicted binding affinity compared to the other samples (Appendix 2).

The results in Figure 5.14 show that there is no correlation between the predicted and the experimental activities for these peptides. Since it has been shown that peptides also bind on the outer surface of lectins (Somers *et al.*, 2000; PDB entry 1G1S), the main problem is probably that these peptides do not bind in a well-defined cavity of the protein. However, our method has been shown to reproduce experimental binding modes for peptides consisting of up to nine amino acids (RMSD values below 1.0 Å between the experimental and docked conformations), provided they bind in a well-defined binding pocket (Paper V). Hence, the development of this method may be a step forward when it comes to docking of flexible ligands such as peptides.

5.4 Design of selective inhibitors of Tyrosine kinase 2

The above examples illustrate some of the limitations with computational docking methods. However, the performance of these methods is highly dependent on the target system, and these methods are most suitable in cases where the ligand binds in a deep pocket or cleft in the protein structure. This is the case for ATP binding to protein kinases. In Paper VI, virtual drug design has been applied to design selective inhibitors that block the binding of ATP to Tyk2.

5.4.1 Method testing and verification of the structural model

To test the influence of the choice of template on the homology models, three alternative homology models were made for the tyrosine kinase domains of Tyk2 and Jak2, using only one template for each model, in addition to homology models made using several templates simultaneously (Paper II). The latter models were assumed to be the most reliable, and were used for drug design. SwissModel (Peitsch, 1995; Peitsch, 1996; Guex and Peitsch, 1997; Guex *et al.*, 1999; Schwede *et al.*, 2003) was used for the homology modelling. Details about the homology modelling of the Tyk2 and Jak2 tyrosine kinase domains can be found in Paper II. Structure superpositioning of the models in Swiss-PdbViewer (Guex and Peitsch, 1997; Swiss-PdbViewer, 2001) gave an RMSD value (average of the pairwise RMSDs) of 1.31 Å between the α-Carbons of the three models of Tyk2, while an RMSD value of 1.18 Å was obtained for the α-Carbons of the three models of Jak2. The Cα RMSD between the model of Tyk2 and the model of Jak2 obtained using five different templates was only 0.75 Å. Hence, the difference between three different models of the same protein made using only one template for each model is larger than the difference between the models of two homologous proteins made using five different templates simultaneously. The relatively high RMSD values between the homology models obtained using a single template indicate that the accuracy of the models is highly dependent on the choice of template. This is a strong argument for using several templates simultaneously in homology modelling. Possibilities for improving the homology model accuracy by combination of several homology models of the same protein are discussed in Paper I.

A set of eleven tyrphostins was docked into the homology model of Jak2, using MOE-Dock (Paper II). One of the tyrphostins was the Jak2 inhibitor AG490, while the other ten tyrphostins were known not to inhibit Jak2. This docking study identified AG490 as the most active compound. This indicates that our homology models are accurate enough to be used for virtual drug design.

For comparison, these eleven tyrphostins were also docked into the homology model of Jak2 using our new gaussian-based docking method presented in Paper V. The same starting conformations as reported in Paper II were used, and the docking analysis was done in the same way as described in Paper VI. The results are given in Table 5.1 (unpublished results).

Table 5.1. Results from docking of the eleven tyrphostins in the homology model of Jak2 using the gaussian-based docking method* (unpublished results).

Tyrphostin	Predicted binding affinity (kJ/mol)
AG490	-38.3
AG1007	-36.9
AG370	-32.2
AG1112	-30.8
AG1478	-26.6
AG294	-24.0
AG126	-20.6
AG18	-19.5
AG30	-18.3
AG879	-3.41
AG1295	-2.90

*The tyrphostins were docked 100 Tabu runs of 1000 iterations each, with a docking box of 3 Å padding around the protein structure.

The results in Table 5.1 show that our docking method is also able to identify AG490 as the most active compound. The estimated binding affinity for AG1007 is however almost as low as for AG490, and AG1007 thus represents a false positive. However, the structures of these two ligands are very similar (see Paper II). MOE-Dock was able to discriminate between these two ligands, but our method uses a much less detailed ligand description than MOE-Dock does. However, the results indicate that our method is suitable for initial screening, where the main purpose is to identify promising compounds. At this stage, it is more important to avoid false negatives than false positives. Though this is a very small testset, the results also indicate that our docking method is suitable for use with homology models, since we were able to identify AG490 as the most active Jak2 inhibitor using a homology model of Jak2.

5.4.2 Database screening and de novo ligand design

The Tyk2 pharmacophore model found by mapping the binding site properties of the homology model of Tyk2 with PASSA (described in Paper II) was used to screen the NCI 3D structure database from August 2000 (http://cactus.nci.nih.gov/) for possible Tyk2 inhibitors (Paper VI). This database contains 250241 structures. The selected molecular fragments from the MCSS (Paper II) defined the pharmacophore (Figure 5.15).

The GROW function of LigBuilder (Wang *et al.*, 2000) was used to design new structures having the proposed functional groups. Structures were built using selected molecular fragments from the MCSS results presented in Paper II as “seed” fragments. The inhibitor binding pocket of Tyk2 was defined by the MCSS fragments that also define the Tyk2 pharmacophore (Figure 5.15). Binding affinities for the resulting structures were estimated using the gaussian-based score function reported in Paper V. Details about the drug design process are given in Paper VI.

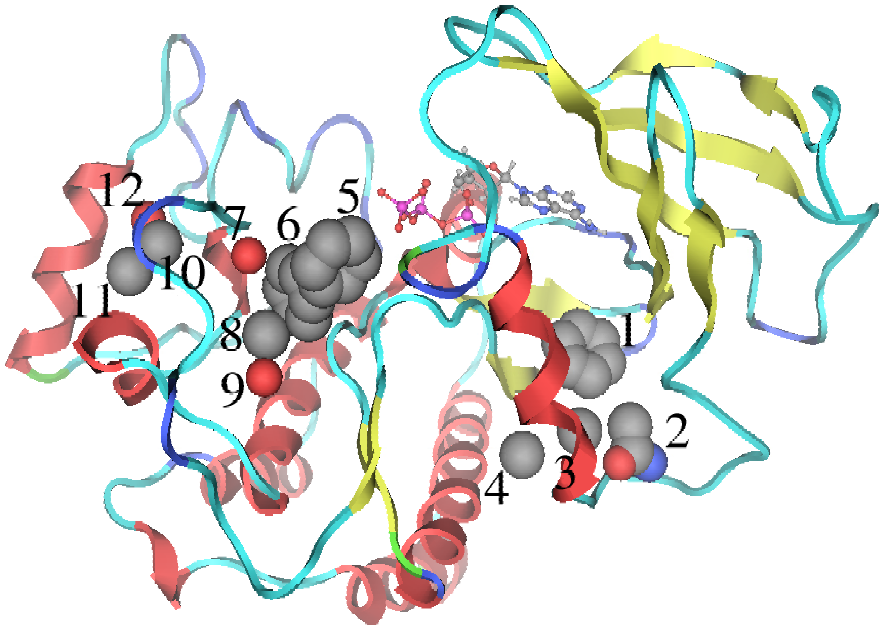


Figure 5.15. The Tyk2 pharmacophore used for the database search and *de novo* ligand design. The pharmacophore was defined by fragments from the MCSS (Paper II).

The hits from the pharmacophore search and the most promising structures from the *de novo* ligand design were docked in the homology model of Tyk2 presented in Paper II, with MOE-Dock (Baxter *et al.*, 1998; MOE, 2002) and with the new gaussian-based docking method introduced in Paper V. In both methods, Tabu search (Baxter *et al.*, 1998) was used for the conformational search. The two docking methods did not identify the same compounds as the most active ones, but they both produced the same conclusion, namely that there are no promising Tyk2 inhibitors in the NCI database. However, our analysis provides useful information about parts of the structures that may be used as functional groups of a selective inhibitor of Tyk2. The main purpose of docking methods is to identify the most active compounds. Most docking methods (as these two) are also trained using X-ray structures of protein-ligand complexes. Hence, internal ranking of inactive compounds is bound to fail, and not interesting for drug design purposes. This may be the reason why the two docking methods ranked the compounds in the NCI database differently. Another explanation might be that the training sets used to derive the score functions in these two methods are different. Treating the receptor as a rigid structure increases the sensitivity to deviations between the system under consideration and the structures in the training set, since ligand-induced conformational changes in the protein structures will have different effects on the results for different types of ligands.

In order to test the promising drug candidates from the pharmacophore search and the *de novo* ligand design for selectivity towards Tyk2, the compounds were docked in the following kinase structures, in addition to the homology models of Tyk2 and Jak2 presented in Paper II: PDB entries 1IR3 (insulin-receptor tyrosine kinase), 1BYG (C-terminal Src kinase), 1FGK (tyrosine kinase domain of fibroblast growth factor receptor 1), 1FPU (Abl kinase), 1QCF (haematopoetic cell kinase, Hck) and 1QPC (lymphocyte-specific kinase, Lck). The gaussian-based docking method was used for this docking study, since it is developed especially for use with homology models. Homology models of both Tyk2 and Jak2 were used here. The results from our docking analysis indicated that none of the structures present in the NCI database can be used to inhibit Tyk2 selectively, but one compound was found to inhibit Tyk2 and insulin receptor tyrosine kinase selectively. However, this study indicated that five of the structures generated by *de novo* ligand design are potential selective inhibitors of Tyk2. The structures of these compounds are shown in Figure 5.16. According to descriptors calculated in MOE, these compounds satisfy Lipinski's "Rule

of five" (Lipinski *et al.*, 1997). The docked structures of these compounds in complex with Tyk2 are shown in Paper VI.

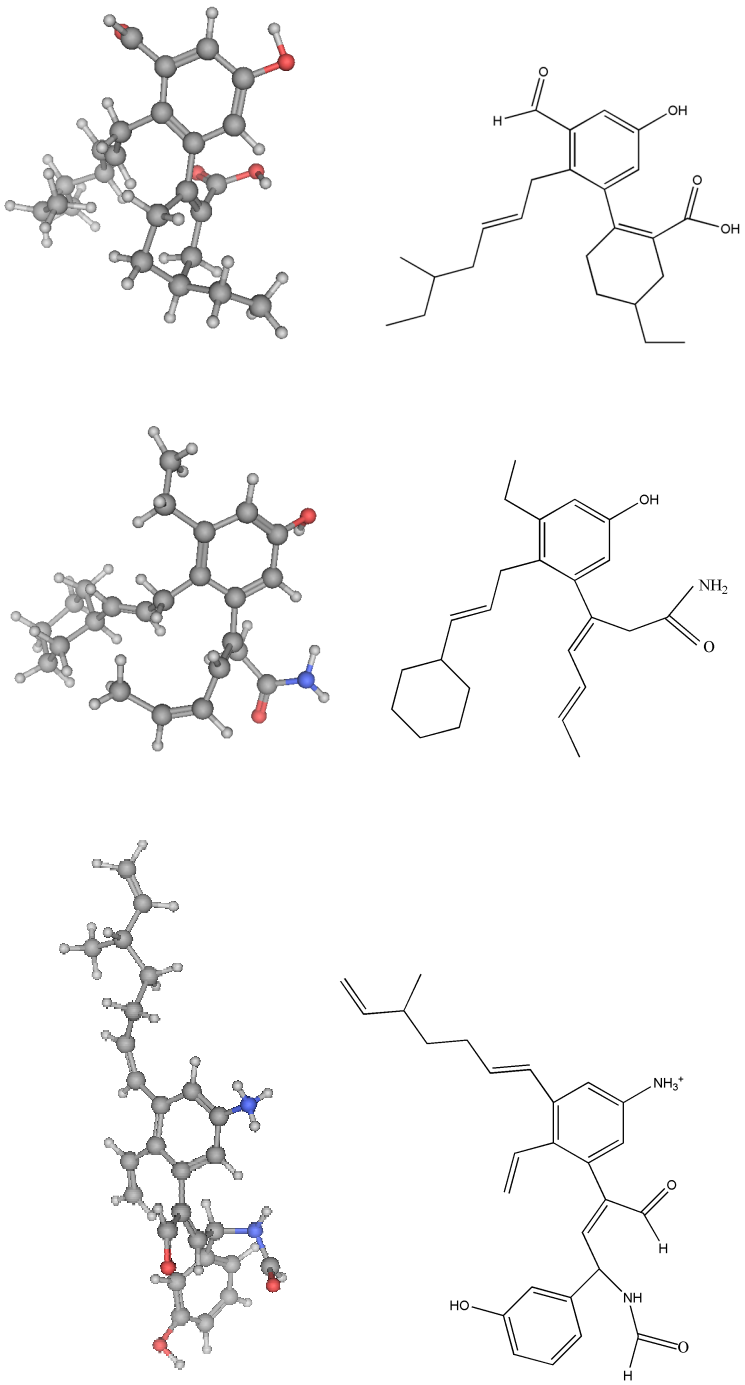


Figure 5.16. The docked conformations of the most promising structures resulting from the *de novo* ligand design.

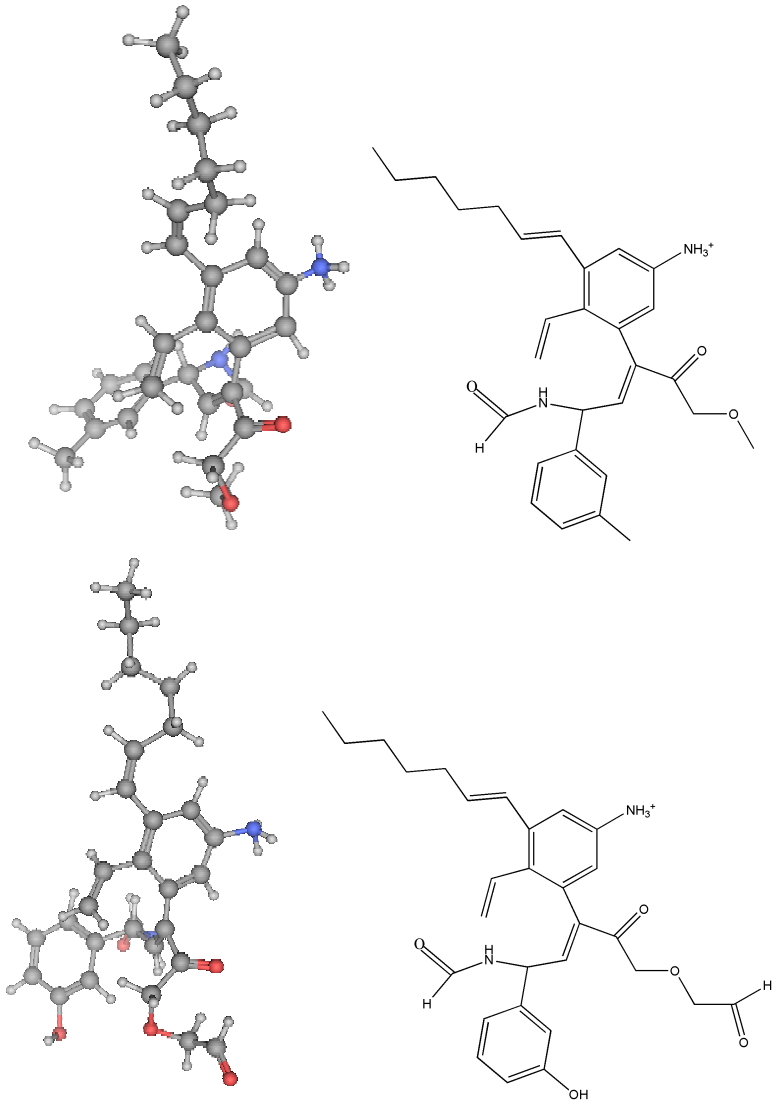


Figure 5.16 (cont.). The docked conformations of the most promising structures resulting from the *de novo* ligand design.

As the results presented in the previous sections clearly demonstrate, the docking methods used in this study have many limitations. However, in this study they are applied to protein kinases, which have a very well-defined binding pocket. Hence, our docking method should be well suited for use with these protein structures. Promising drug candidates could have been missed due to insufficient representation of hydrogen bond formation. However, the fact that neither MOE-Dock nor our gaussian-based docking method identified any promising Tyk2 inhibitors in the NCI database indicates that we can trust the results. Many hydrophobic groups are present among the functional groups proposed for a selective Tyk2 inhibitor. This indicates that hydrophobic interactions are important for binding of an inhibitor in this binding pocket. This gives additional confidence in the results since our docking method has been shown to predict hydrophobic interactions quite well (Paper V).

5.5 Application of Protein Alpha Shape Similarity Analysis (PASSA) in modelling selectivity

When promising compounds have been identified, it is important to test the selectivity by predicting their affinity to a number of related proteins, in addition to the target. In Paper VII, the binding sites of two sets of protein structures were mapped using PASSA, and empirical models were trained by relating the structural properties of the protein binding sites to the affinity of different ligands towards the proteins. These empirical models can be used to predict the affinity of the ligands towards related proteins based on the gaussian property fields for the protein binding sites. This is useful for detection of possible side effects of the drug candidates. 3D-QSAR methods predict the activities of new ligands towards a protein using a regression model that relates the structural properties of known ligands to their experimental affinities. Here we relate the structural properties of the protein binding sites of several proteins to the activities of ligands towards these proteins. The activities of these ligands towards related proteins can then be predicted using the obtained regression model.

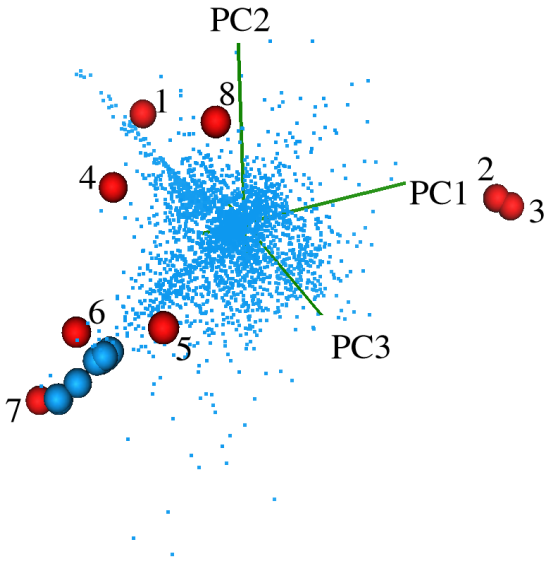
In Paper II, it was shown that with PASSA, interactions known to be important for the selectivity of STI-571 towards Abl kinase could be identified. This indicates that PASSA is a useful tool for modelling selectivity, making it a useful supplement to virtual screening with computational docking. Empirical docking methods are trained on diverse sets of compounds and are meant to be as general as possible. Hence, the results for a certain protein-ligand complex depend on the similarity of the complex to the structures used to train the method. Using PASSA to model selectivity within a protein family, as in this work, allows for more detailed and family-specific modelling of protein-ligand interactions. This method also allows for effective visualisation of the molecular basis for selectivity.

In Paper VII, PASSA has been used to model selectivity of ligands towards two sets of protein kinase structures. Dataset 1 contains a set of eight protein kinase C (PKC) isozymes (Jirousek *et al.*, 1996), while dataset 2 consists of a set of structures of the kinase domains of Abl kinase, epidermal growth factor receptor (EGFR), platelet-derived growth factor receptor (PDGFR), c-Src, protein kinase A (PKA) and two isozymes of PKC (Zimmermann *et al.*, 1997).

The PLS regression used to model the experimental IC_{50} values produces loading weights and regression coefficient vectors containing elements for every grid point. Hence, they can be mapped back onto the grid points used to compute the gaussian property fields, and structural regions that contribute to selectivity may be identified. By selecting interesting regions in a plot of the loading weights, the corresponding interaction sites in the proteins can be identified. An example of this is given in Figure 5.17, where the selected loading weights from the regression model made for the PKC isozymes (dataset 1) are shown in the upper part of the figure, while the corresponding regions in the protein binding site are shown below.

As illustrated in Figure 5.17 A, one interesting set of loading weights is protruding towards proteins ‘1‘ (PKC-α) and ‘4’ (PKC-γ), and another towards protein ‘7’ (PKC-ζ). Figure 5.17 B shows that the loading weights spanning the direction of protein ‘7’ correspond to a well-defined hydrophobic region in the structure of PKC-ζ.

A:



B:

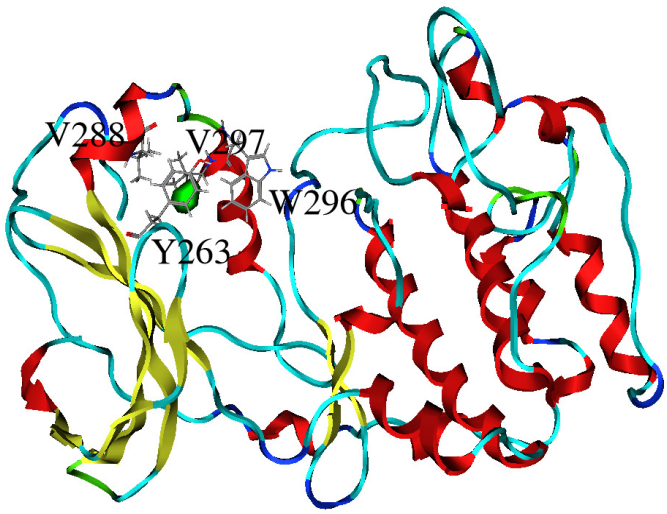


Figure 5.17. **A:** Scores and loading weight bi-plot. The PLS scores of each protein are shown as red spheres and the PLS loading weights of the grid variables are shown as blue dots. The selected loading weights are rendered as blue spheres. **B:** The structural origin (in PKC-ζ) of the selected loading weights. Hydrophobic sites are shown in green.

Since experimental structures are available for several of the protein-ligand complexes in dataset 2, we are able to test whether the results produced by our method correspond to the structural properties of the actual ligands. This can be used as a test on how well the results from PASSA overlap with the properties of known, selective inhibitors. In Figure 5.18, the regression coefficients for the selective Abl kinase inhibitor STI-571 are plotted together with the X-ray structure of Abl kinase in complex with this inhibitor (PDB entry 1IEP).

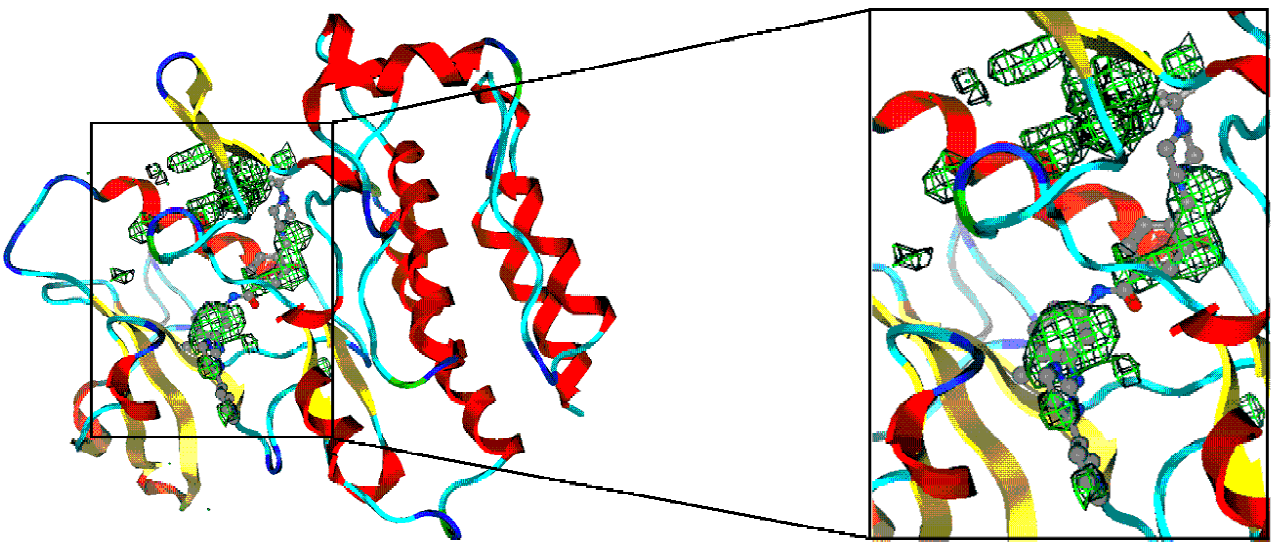


Figure 5.18. The regression coefficients for the hydrophobicity (green) for STI-571 plotted together with the X-ray structure of Abl kinase in complex with STI-571 (PDB entry 1IEP). STI-571 is rendered in "ball and stick".

The results in Figure 5.18 show that the regression coefficients for the hydrophobicity for STI-571 correspond well with the positions of the hydrophobic groups of STI-571. This indicates that PASSA is a useful method for identification of regions in a protein binding site that can be utilised to achieve selective binding of ligands to the protein. The results presented in Paper VII also demonstrate that the PASSA method may be used quantitatively to predict IC_{50} values for a number of ligands towards a set of closely related protein targets. This makes PASSA a promising method in screening for side effects.

6 Discussion

With the large amount of data resulting from the human genome project (McPherson *et al.*, 2001), homology modelling of protein structures has received increasing attention in drug design. However, this method uses many approximations, and improvements of existing methods are required for reliable results. A comparison of three alternative homology models of Tyk2 and Jak2 showed that the deviation between three different models of the same protein made using a single template for each model was larger than the deviation between the models of Tyk2 and Jak2 made using five different templates simultaneously. This indicates that the accuracy of the models is highly dependent on the choice of template, and is a strong argument for using several templates simultaneously in homology modelling.

Most drug design methods are trained on more accurate protein structure models resulting from X-ray crystallography and NMR experiments. Hence, few methods exist that are suitable for use with homology modelled protein structures. The development of new computational methods that can deal with small structural inaccuracies like those found in homology models is therefore becoming more and more important. Methods that ignore receptor flexibility are especially sensitive to errors in the protein structure models. The work presented in this thesis has focused on development of methods utilising gaussian functions to describe molecular properties. These methods are regarded to be more suited for use with homology modelled protein structures than e.g. force field based methods. The gaussian functions introduce a smoothing of the molecular surface descriptions, which decreases the sensitivity to errors in the structural models used. The results presented here indicate that drug design methods utilising gaussian functions to describe molecular properties have many applications and a great potential in structure-based drug design. These methods are relatively fast, and well suited for the initial stages of a drug design process, when the goal is to identify the main features of ligand binding. However, the results produced with these methods may be less accurate than results obtained using more time consuming methods that use more variables and detailed information to represent the protein-ligand interactions. The smooth molecular description used by our docking method might also lead to an overestimation of the binding affinity for ligands having many similarities to the real drug lead. However, the results indicate that our method is suitable for initial screening, where the purpose is to identify promising compounds, and false negatives are more important to avoid than false positives. Even the most accurate methods use crude approximations to the real system, and can not always represent the many factors involved in the complex process of ligand binding. The results from a computational docking study using flexible carbohydrates and peptide ligands binding to the outer surface of E-selectin clearly illustrate the many limitations these methods have. Hence, in spite of the many applications and the proven usefulness of docking methods in pharmaceutical research, it is important to realise their limitations. It is important to compare the results produced by several different methods, since otherwise one can easily be misled, and important drug leads can be missed.

In general, computational docking methods perform best on small ligands with few rotatable bonds. Compounds such as large carbohydrates and peptides, and especially those that bind to the outer surface of the protein are difficult to model. The main reason is that most empirical score functions are not trained on this class of compounds, but on smaller ligands binding in well-defined cavities in the protein structure. This is probably caused by the bias in the datasets available for training score functions. The same is true for the knowledge-based score functions, although the bias is lower here, since the development of these functions is not dependent on binding affinity data to derive the rules. The methods used to describe the protein binding site properties may also be more suitable for well-defined binding pockets than for the highly hydrophilic outer surface of the protein. Force field based score functions suffer from being time consuming, but may give better results for example for ligands binding to the outer protein surface, since they are not affected

by the bias in the available X-ray structures. However, as the number of rotatable bonds increases, the computational time required to obtain reliable results also increases enormously. Many force field based methods substitute the free energy of binding in solution by an estimate of the gas-phase enthalpy of binding. Hence, both solvent and entropic effects are ignored. Getting trapped in local minima is also a common problem with these methods. Methods based on gaussian functions have advantages when it comes to speed and robustness against small structural errors, but they are most suitable for the initial screening stages of a drug design process. These methods have for example been shown to model electrostatics and hydrogen bond formation insufficiently. For reliable prediction of binding affinities and active conformations, more accurate methods have to be applied. Hence, a virtual drug design process is most likely to succeed if a combination of several methods with different levels of accuracy is used.

Due to induced fit, development of methods that include protein flexibility in the calculations is important. To date, few computationally efficient methods have been developed to handle this problem, but this is one of the major research areas in rational drug design. In spite of their many limitations, computational drug design methods have made important contributions to pharmaceutical research, and given a suitable model system, and when used with some critical sense, they are effective tools that speed up the drug design process.

7 Conclusions

In this work, a new method for prediction of homology model accuracy with multivariate regression has been developed. This method predicts the model accuracy directly from the amino acid sequence alignment and can be used to assure that the optimal templates and alignments are utilised, so that the best possible homology model is generated. It is also useful for identification of structural regions that are difficult to model, as well as errors in the sequence alignment. Here, the method has been applied to the protein kinase family, but it can easily be extended to other protein families.

A gaussian-based method for mapping protein binding site properties and identification of possible interaction sites for selective inhibitors, and a gaussian-based computational docking method have also been developed. These methods have been shown to be fast, and suitable for virtual screening. The gaussian-based docking method runs ten times as fast as for example MOE-Dock, and has been shown to perform well on relatively large and flexible ligands, such as peptides, provided they bind in a well-defined binding pocket. This method was able to reproduce the experimental conformations of peptides containing up to nine amino acids. Since we include no information about hydrogen atom positions and partial charges, our score function is unable to represent direction-specific hydrophilic interactions and formation of ion bonds between the protein and the ligand.

PASSA, the gaussian-based method for mapping protein binding sites presented here, has been tested on protein kinases bound to known, selective inhibitors. The results indicate that ligand properties and interaction sites in the protein binding pocket that are important for selectivity can be identified with PASSA. PASSA and the gaussian-based docking method developed here have been utilised in a rational drug design process, resulting in suggested structures for five selective Tyk2 inhibitors and one inhibitor of Tyk2 and insulin receptor tyrosine kinase. In this work, PASSA and the gaussian-based docking method were combined with database screening, MCSS and *de novo* ligand design. PASSA has also been used to model the activities of a number of ligands towards protein kinases, and has been shown to be a promising method in screening for side effects.

In addition to the design of Tyk2 inhibitors, the interactions between the receptor kinase FGFR1 and a known inhibitor have been studied, and several improvements of this inhibitor have been suggested by computational sensitivity analysis and comparative database analysis.

8 Future perspectives

Our gaussian-based score function predicts hydrophobic interactions better than hydrophilic interactions, while many other score functions for computational docking predict hydrophilic interactions better than hydrophobic interactions (Wang *et al.*, 2003). Hence, it would be interesting to combine our score function with other score functions to improve the predictive ability. We plan to develop a new version of the gaussian-based docking method that can better represent hydrogen bonds and ion interactions. Until now, the main focus has been on method development. Hence, the performance of the docking method has to be verified further, and compared to other docking methods. We have only tested our gaussian-based score function using Tabu search for the conformational searching. It would be interesting to test the performance of the score function using a different conformational search algorithm as well, for example a genetic algorithm. We also plan to develop new versions of both PASSA and the gaussian-based docking method that are independent of commercial software packages, in order to make these methods more available to the research community. It would also be interesting to test the robustness of our new gaussian-based docking method against small structural errors, by docking ligands of known activity into homology models of the target proteins. We also plan to test the robustness by perturbing the X-ray structures of the proteins in the training set, and docking the ligands into ensembles of protein structure models.

In the work presented here, five candidate structures for a selective Tyk2 inhibitor were suggested. We plan to test the activity of these compounds further, in a cell assay.

PASSA has been shown to identify properties corresponding to active groups of inhibitors that are known to be important for the selectivity of these inhibitors towards their target proteins. Hence, an idea would be to develop a *de novo* ligand design program based on gaussian property descriptions for both the protein and for molecular fragments from e.g. MCSS. Gaussian property fields could then be generated for different combinations of these fragments. One could also think of a method using the growing-approach to design the ligand, where ligand groups are added incrementally, and new ligand property fields are generated and compared to the receptor fields. It would also be interesting to test the performance of PASSA further, on all PDB entries holding selective protein inhibitors.

PASSA combined with DPLSR may be a suitable method for diminishing the dependency of homology models upon the applied template structures, since this approach has been shown to be able to single out unique properties of proteins. The residuals from DPLSR using ensembles of homology models of each protein might contain information about properties that are common to all homology models of a given protein. The dependent variables in the DPLSR should then be indicator variables indicating the template structures used for the homology modelling.

In general, including solvent effects and protein flexibility efficiently, and increasing the suitability of the methods for homology modelled proteins are the major challenges in the development of new rational drug design methods to date. Accomplishing this will increase both the hit-rate and the number of targets that can be considered significantly.

Appendix 1. Computational details of the docking of carbohydrate ligands and peptides in E-selectin

The crystal structure of the lectin and epidermal growth factor (EGF)-like domains of E-selectin was obtained from PDB entry 1ESL. Hydrogen atoms were added to the X-ray structure in MOE, and the structure was energy minimised with AMBER94 (Weiner *et al.*, 1984) until convergence with an RMSD gradient of 0.1.

Since the conformation of $SiaLe^x$ has been shown to change upon binding to E-selectin (Cooke *et al.*, 1994), the bioactive conformation was manually reproduced based on experimental data for atom distances and dihedral angles (Scheffler *et al.*, 1995). This conformation was used as starting conformation in the docking study. Starting conformations for the remaining carbohydrate ligands in the set (Vodovozova *et al.*, 2000) were generated by manual alignment to $SiaLe^x$. Two different docking methods were used: docking by simulated annealing in MOE (Hart and Read, 1992), and our gaussian-based docking method presented in Paper V. Both docking calculations were performed using MMFF94 (Halgren, 1996) with a smooth non-bonded cut-off of 10-12 Å.

Each simulated annealing run consists of a sequence of Monte Carlo cycles, each consisting of a number of random changes of the atom coordinates (Hart and Read, 1992). The temperature is held constant during each cycle, and is systematically reduced from one cycle to the next. Each cycle after the first cycle begins with the lowest energy conformation from the previous cycle. Each cycle continues until either the number of accepted changes or the number of rejected changes reaches the iteration limit. The initial simulated temperature (the simulated temperature maintained during the first cycle of each run) was 1000 K. For each carbohydrate ligand, 80 docking runs of 30 cycles each were performed. The iteration limit was 8000, and a docking box with 2 Å padding around the set of aligned carbohydrate ligands was used.

Both the carbohydrate ligands and 25 E-selectin binding peptides (Martens *et al.*, 1995) were docked 100 Tabu runs of 1000 iterations each in the X-ray structure of E-selectin, using our gaussian-based docking method. The docking analysis of the carbohydrates was carried out in the same way as described in Paper VI, using the starting conformations described above. The peptides were docked from a random starting conformation, using a docking box with 3 Å padding around the receptor and the aligned structures of all peptides. The AMBER94 force field (Weiner *et al.*, 1984) was used for the peptides.

Appendix 2. Results from the docking of carbohydrate and peptide ligands in E-selectin

The docking results and experimental binding affinity data for the carbohydrate and peptide ligands are given in Table A2.1 and Table A2.2, respectively.

Table A2.1. Docking results and experimental binding affinity data for the carbohydrate ligands.

	Carbohydrate ligand	Experimental binding affinity (0-5)*	Docking energy from MOE-Dock (kcal/mol)	Predicted binding affinity (kcal/mol)
1	3'-HSO_3Le^x	1	59.8	0
2	α-Neu5Ac	1	-177.8	-0.33
3	α-L-Fuc	1	-94.2	-0.61
4	α-L-Rha	1	-83.2	-0.55
5	Le^a-trisaccharide	2-3	66.7	-0.50
6	3'-HSO_3Le^a	1-2	-62.5	-0.51
7	α-D-Man	0-1	-86.7	-0.61
8	α-D-Man-6-phosphate	0-1	-333.5	-0.51
9	A-trisaccharide	3	-7.44	0
10	6-Sia-Lac	1-2	5.81	-0.38
11	$(Neu5Ac\alpha2\text{-}8)_3$	1	29.3	0
12	Galα1-3GalNAcα	0-1	-6.68	-0.077
13	$(Neu5Ac\alpha2\text{-}8)_2$	1	-94.3	-0.19
14	α-D-glucose	0	-89.6	-0.61
15	Neu5Acα2-3Galβ1-4Glc	1	-0.68	-0.21
16	$SiaLe^a$	2	249.9	0
17	$SiaLe^x$	5	309.4	0
18	Le^d (H type 1)	2	1.05	0
19	β-D-glucose	0	-102.2	-0.60
20	B_{di}	2	-61.0	0
21	β-GlcNAc	1	-128.6	0
22	β-D-galactose	1	-106.7	-0.58
23	A_{di}	2-3	-85.6	0
24	B-trisaccharide	3	17.5	0
25	Le^x	2	-20.2	-0.42
26	Le^y	1	44.9	-0.56
27	β-GalNAc	1	-132.0	0
28	α-GalNAc	1	-121.8	-0.81
29	β-D-GlcNAc-6-sulfate	0	-196.5	-0.21
30	Galβ1-3GalNAc	1	-107.7	-0.39
31	$6HSO_3LactNAc$	1	-100.4	-0.010
32	HSO_3Gal	1	-183.6	-0.55
33	Neu5Acα2-6GalNAcα	1	-62.9	-0.025
34	GlcNAcβ1-4GlcNAc	0-1	-79.3	-0.40

*The affinity of the carbohydrates to the cancer cells was estimated by microscopic evaluation of fluorescence emitted from the bound carbohydrate ligands (Vodovozova *et al.*, 2000). In cases where the experimental binding affinity lies between two values, the lowest value was chosen. The affinity for $SiaLe^x$ was originally given as "Very bright fluorescence". We assumed that this corresponds to the value 5, compared to the other samples.

Table A2.2. Docking results and experimental binding affinity data for the E-selectin binding peptides.

Nr.	Peptide	Experimental IC_{50} (nM)*	Predicted binding affinity (kcal/mol)
1	H_2N-TWDQLWDLMK-COOH	2500	-1.20
2	H_2N-ITWDQLWDLMK-COOH	11	-1.49
3	H_2N-DITWDQLWDLMK-COOH	4	-1.63
4	H_2N-TWDQLWDLMK-$CONH_2$	4	-1.06
5	Ac-TWDQLWDLMK-$CONH_2$	9	-2.26
6	H_2N-DITWDQLWDLM-COOH	460	-1.77
7	H_2N-DITWDQLWDL-COOH	150000	-2.02
8	H_2N-DYTWFELWDMMQ-COOH	11	-1.92
9	H_2N-DITWDELWKIMN-COOH	4.4	-1.71
10	H_2N-DYSWHDLWEMMS-COOH	57	-1.14
11	H_2N-QITWAQLWNMMK-COOH	16	-2.17
12	H_2N-HITWDQLWRIMT-COOH	83	-1.67
13	H_2N-HVSWEQLWDIMN-COOH	76	-1.15
14	H_2N-DMTWHDLWTLMS-COOH	23	-0.040
15	H_2N-EITWDQLWEVMN-COOH	67	-1.24
16	H_2N-DISWDDLWIMMN-COOH	620	-1.32
17	H_2N-QITWDQLWDLMY-COOH	910	-1.44
18	H_2N-HRAEWLALWEQMSP-COOH	47	-0.061
19	H_2N-KKEDWLALWRIMSV-COOH	71	-0.73
20	H_2N-RNMSWLELWEHMK-COOH	5.4	-1.32
21	H_2N-AEWTWDQLWHVMNPAESQ-COOH	23	-2.87
22	H_2N-KRKQWIELWNIMS-COOH	1200	-2.22
23	Ac-WKLDTLDMIWQD-$CONH_2$	>30000	-0.038
24	H_2N-HITWDQLWNVMN-COOH	420	-1.60
25	H_2N-HITWDQLWNVMLRRASLG-COOH	>11000	240.2

*Data from (Martens *et al.*, 1995).

Appendix 3. Multiple sequence alignment and alignment score profiles for the kinases studied in Paper I

The multiple sequence alignment and alignment score profiles for alignment position 1-150, 150-300 and 300-450 are shown in Figure A3.1, A3.2 and A3.3, respectively (data from Figure 4 in Paper I).

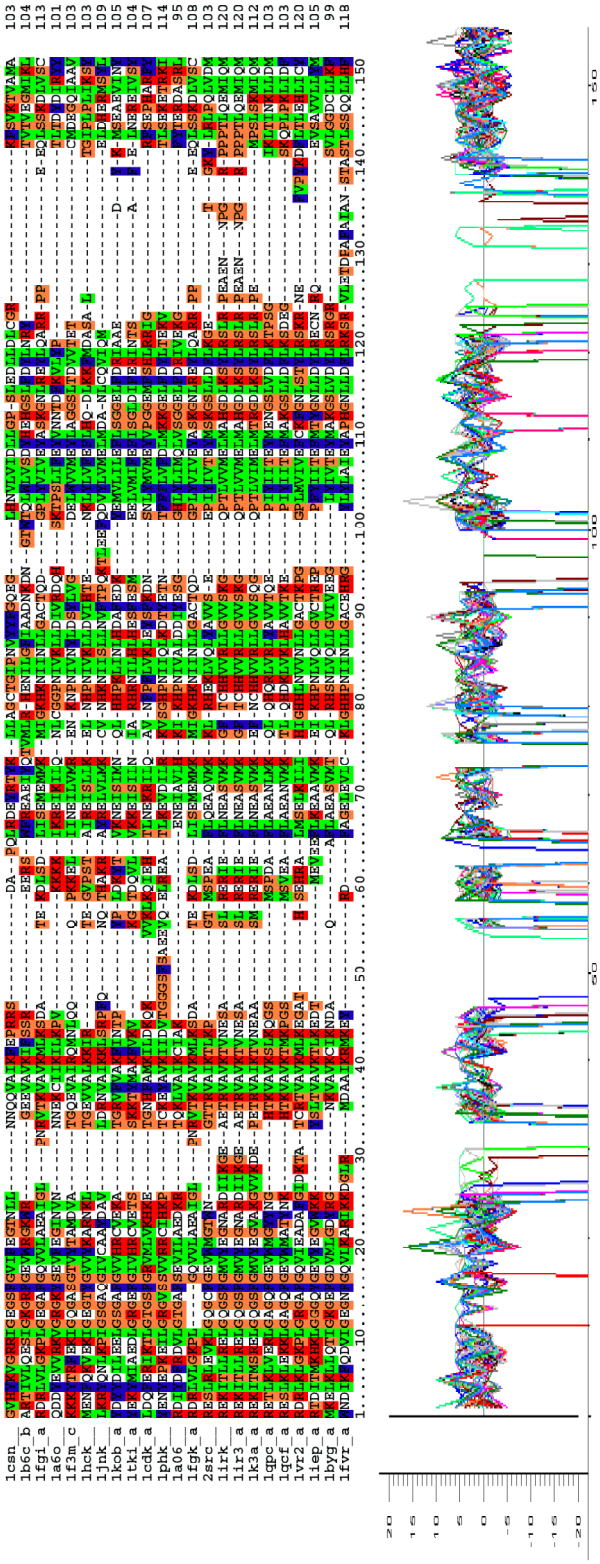


Figure A3.1. Multiple sequence alignment and alignment score profiles for alignment position 1-150 (data from Figure 4 in Paper I).

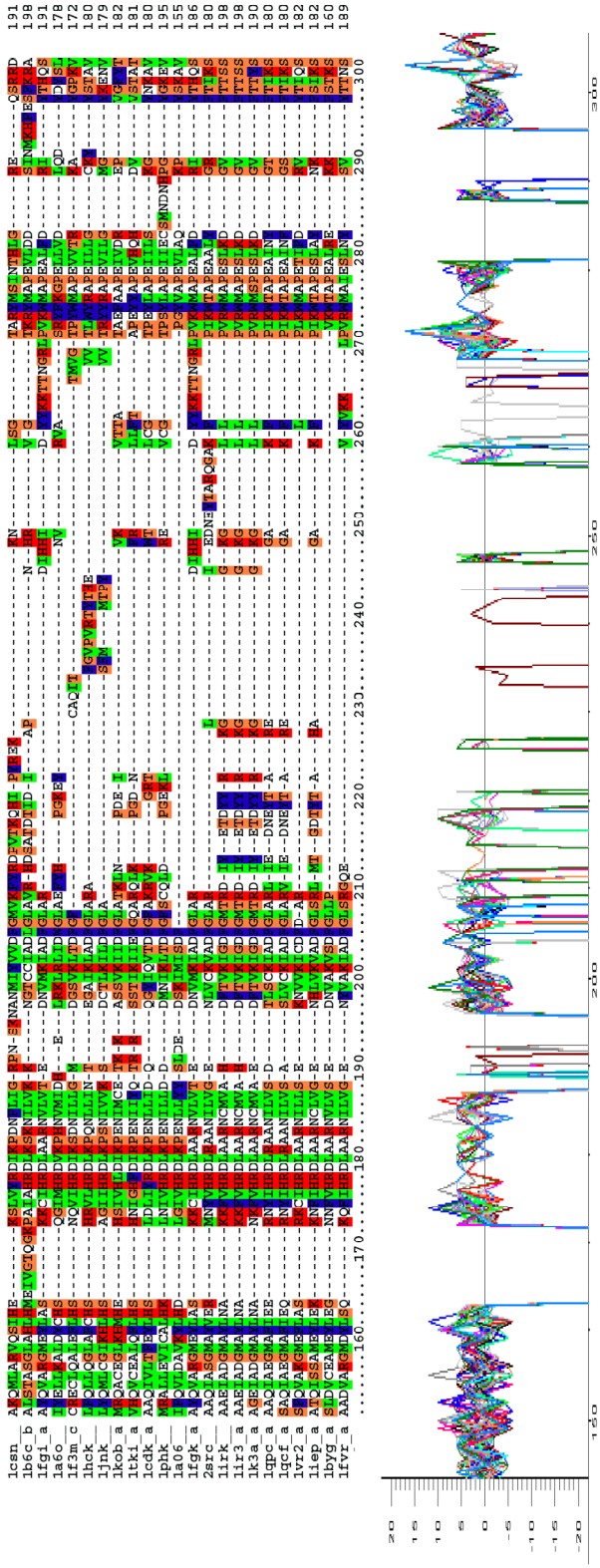


Figure A3.2. Multiple sequence alignment and alignment score profiles for alignment position 150-300 (data from Figure 4 in Paper I).

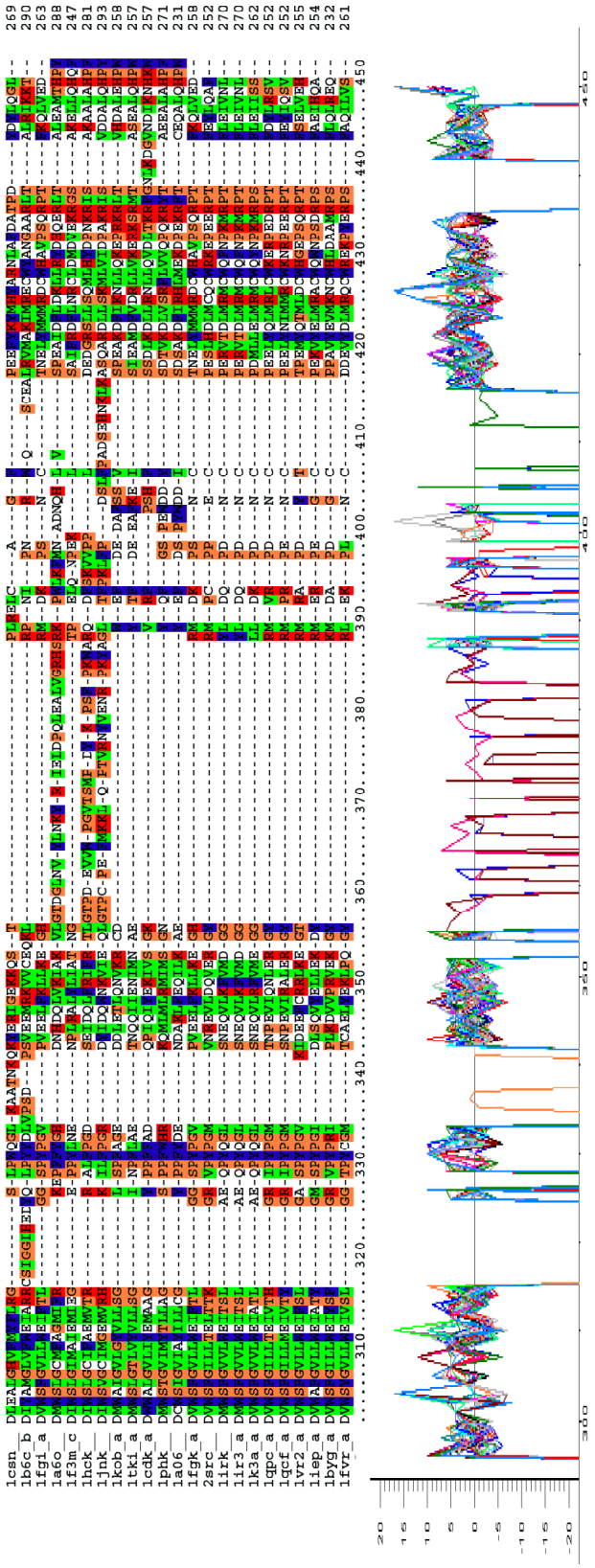


Figure A3.3. Multiple sequence alignment and alignment score profiles for alignment position 300-450 (data from Figure 4 in Paper I).

Appendix 4. Multiple sequence alignment and regression coefficients from Paper I

The multiple sequence alignment and regression coefficients for alignment position 1-150, 150-300 and 300-450 are shown in Figure A4.1, A4.2 and A4.3, respectively (data from Figure 8 in Paper I).

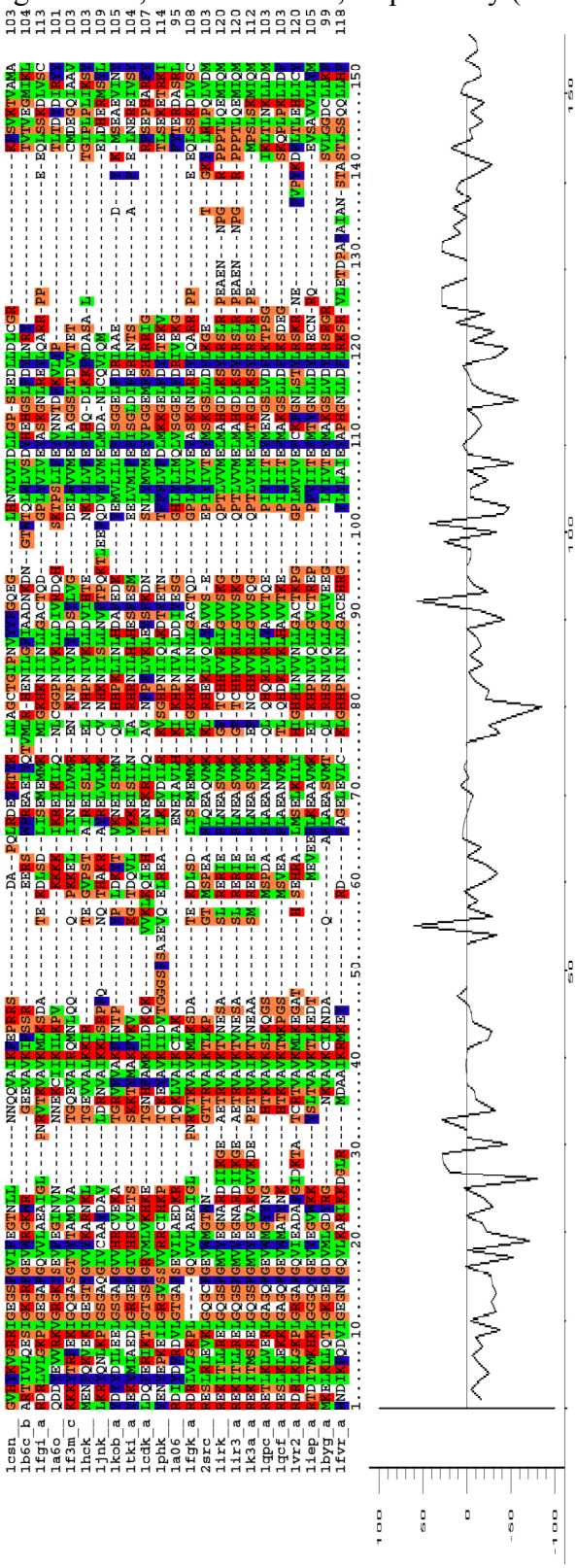


Figure A4.1. Multiple sequence alignment and regression coefficients for alignment position 1-150 (data from Figure 8 in Paper I).

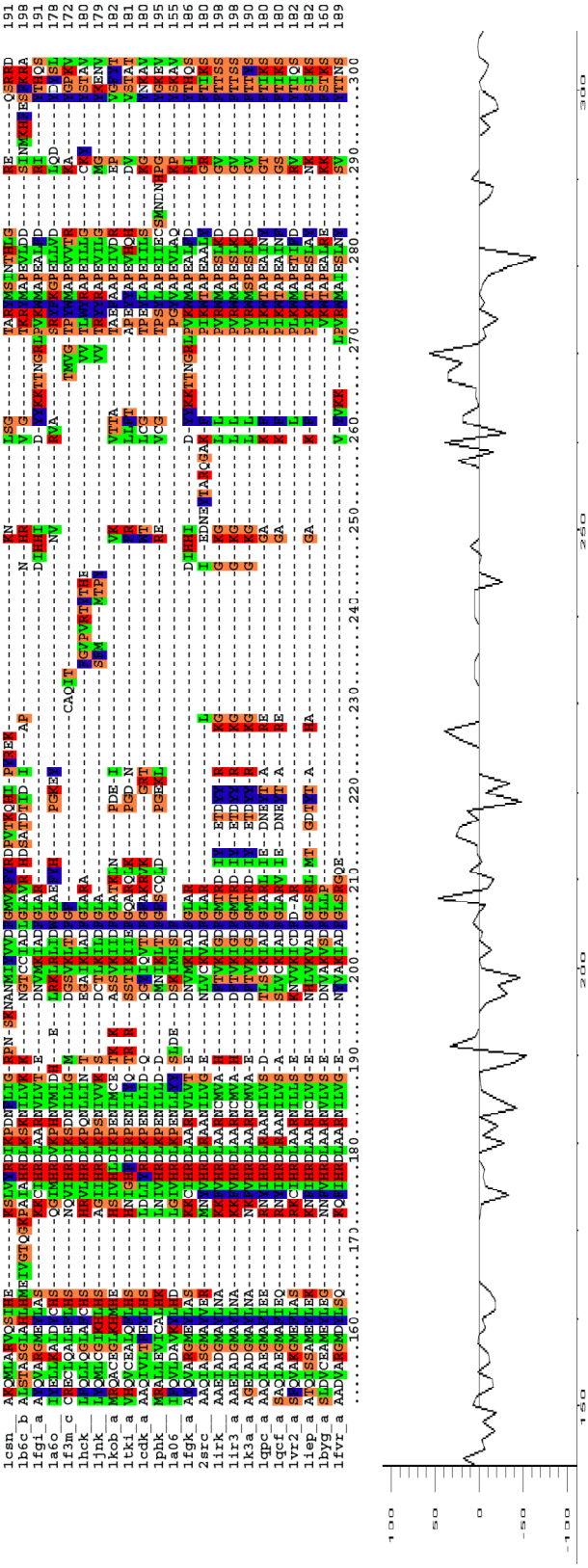


Figure A4.2. Multiple sequence alignment and regression coefficients for alignment position 150-300 (data from Figure 8 in Paper I).

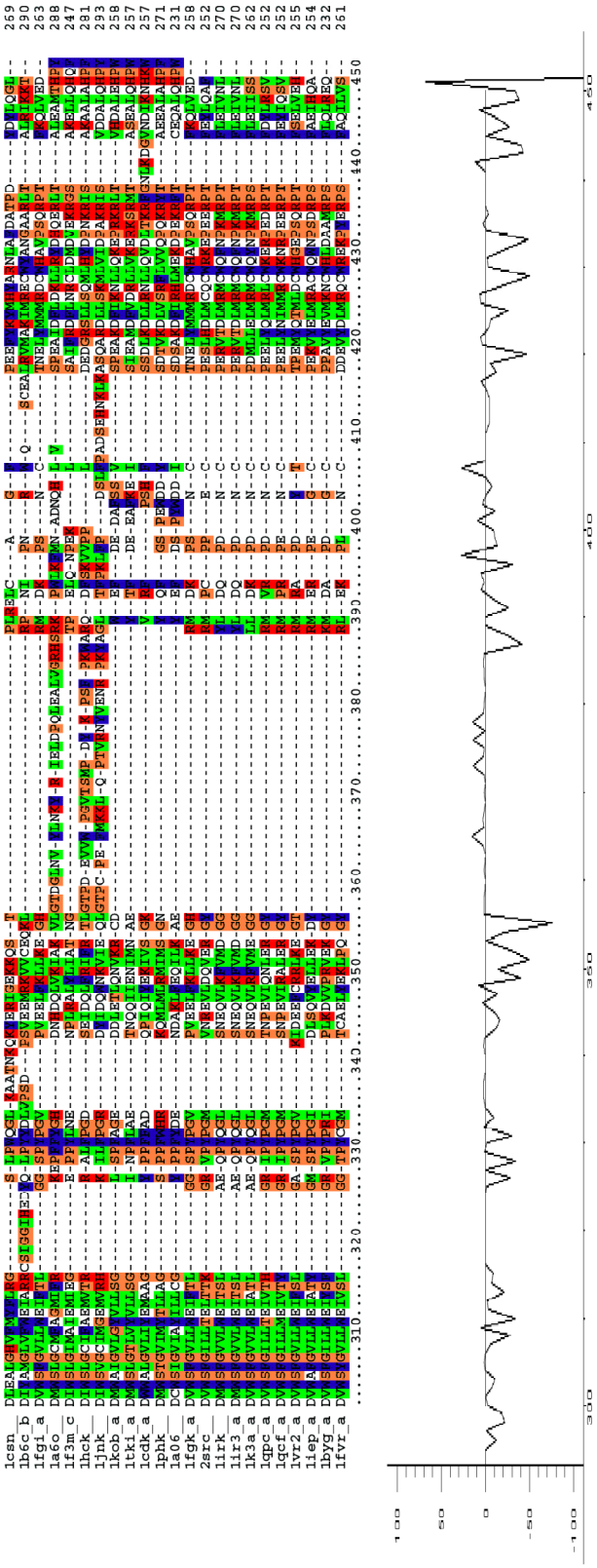


Figure A4.3. Multiple sequence alignment and regression coefficients for alignment position 300-450 (data from Figure 8 in Paper I).

References

Abagyan, R. A.; Totrov, M. M. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **1997**, *268*, 678-685.

Al-Lazikani, B.; Jung, J.; Xiang, Z.; Honig, B. Protein structure prediction *Curr. Opin. Chem. Biol*. **2001**, *5*, 51-56.

Allen, F. H.; Kennard, O.; Taylor, R. Systematic analysis of structural data as a research technique in organic-chemistry. *Acc. Chem. Res.* **1983**, *16*, 146-153.

Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford University Press: New York, 1989.

Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787-797.

Anderson, A. C.; O'Neil, R. H.; Surti, T. S.; Stroud, R. M. Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chem. Biol.* **2001**, *8*, 445-457.

Anderson, K. Advances in the biology of multiple myeloma: Therapeutic applications. *Semin. Oncol.* **1999**, *26*, 10-22.

Apostolakis, J.; Caflisch, A. Computational ligand design. *Comb. Chem. High. T. Scr.* **1999**, *2*, 91-104.

Bajorath, J. Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* **2001**, *6*, 989-995.

Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882-894.

Bajorath, J.; Stenkap, R.; Aruffo, A. Knowledge-based model-building of proteins- concepts and examples. *Protein Sci.* **1993**, *2*, 1798-1810.

Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93-96.

Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Prot. Struct. Func. Gen.* **1998**, *33*, 367-382.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28*, 235-242.

Böhm, H.-J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided Mol. Des.* **1992**, *6*, 593-606. (a)

Böhm, H.-J. The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* **1992**, *6*, 61-78. (b)

Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. *J. Comput. Aided Mol. Des.* **1994,** *8***,** 243-256.

Böhm, H.-J.; Stahl, M. The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2002; 18, pp 41-87.

Böhm, M.; Stürzebecher, J.; Klebe, G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458-477.

Brady, G. P.; Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **2000**, *14*, 383-401.

Branden, C.; Tooze, J. *Introduction to protein structure*, 2nd ed.; Garland Publishing, Inc.: New York, 1999.

Bright, J. J.; Du, C. G.; Sriram, S. Tyrphostin b42 inhibits IL-12-induced tyrosine phosphorylation and activation of Janus kinase-2 and prevents experimental allergic encephalomyelitis. *J. Immunol.* **1999**, *162*, 6255-6262.

Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335-373.

Broughton, H. B. A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening. *J. Mol. Graph. Model.* **2000**, *18*, 247-257.

Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. ISOSTAR: a library of information about non-bonded interactions. J. *Comput. Aided Mol. Des.* **1997**, *11*, 525-537.

Burley, S. K.; Almo, S. C.; Bonanno, J. B.; Capel, M.; Chance, M. R.; Gaasterland, T.; Lin, D. W.; Sali, A.; Studier, F. W.; Swaminathan, S. Structural genomics: beyond the Human Genome Project. *Nat. Genet.* **1999**, *23*, 151-157.

Capdeville, R.; Buchdunger, E.; Zimmermann, J.; Matter, A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat. Rev. Drug Discov.* **2002**, *1*, 493-502.

Cardozo, T.; Batalov, S.; Abagyan, R. Estimating local backbone structural deviation in homology models. *Comput. Chem.* **2000**, *24*, 13-31.

Carlson, H. A. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447-452.

Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. Method for including the dynamic fluctuations of a protein in computer-aided drug design. *J. Phys. Chem. A* **1999**, *103*, 10213-10219.

Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **2000** *57*, 213-218.

Carter-Su, C.; Smit, L.S. Signaling via JAK tyrosine kinases: Growth hormone receptor as a model system *Recent. Prog. Horm. Res.* **1998**, *53*, 61-83.

Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209-225.

Chan, H. S.; Dill, K. A. The protein folding problem. *Phys. Today* **1993**, *46*, 24-32.

Chinea, G.; Padron, G.; Hooft, R. W.; Sander, C.; Vriend, G. The use of position-specific rotamers in model building by homology. *Prot. Struct. Func. Gen.* **1995**, *23*, 415-421.

Chothia, C.; Lesk, A. M. Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin. *J. Mol. Biol.* **1982**, *160*, 309-323.

Chothia, C.; Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823-826.

Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377-395.

Cline, M.; Hughey, R.; Karplus, K. Predicting reliable regions in protein sequence alignments. *Bioinformatics* **2002**, *18*, 306-314.

Cohen, F. E.; Sternberg, M. J. E.; Taylor, W. R. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **1982**, *156*, 821-862.

Connolly, M. L. Analytical molecuar surface calculation. *J. Appl. Cryst.* **1983**, *16*, 548-558.

Cooke, R. M.; Hale, R. S.; Lister, S. G.; Shah, G.; Weir, M. P. The conformation of the sialyl Lewis x ligand changes upon binding to E-selectin. *Biochemistry* **1994**, *33*, 10591-10596.

Cramer, C. J. *Essentials of computational chemistry. Theories and models*; John Wiley & Sons, Ltd.: Chichester, 2002.

Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular-field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.

Creighton, T. E. *Proteins. Structures and molecular properties*, 2nd ed.; W. H. Freeman and Company: New York, 1993.

Cristobal, S.; Zemla, A.; Fischer, D.; Rychlewski, L.; Elofsson, A. A study of quality measures for protein threading models. *BMC Bioinformatics* **2001**, *2*, *5*.

Dahl, S. G.; Kristiansen, K.; Sylte, I. Bioinformatics: from genome to drug targets. *Ann. Med.* **2002**, *34*, 306-312.

David, A.; Kopeckova, P.; Minko, T.; Rubinstein, A.; Kopecek, J. Design of a multivalent galactoside ligand for selective targeting of HPMA copolymer-doxorubicin conjugates to human colon cancer cells. *Eur. J. Cancer* **2004**, *40*, 148-157.

Davis, M. E.; Madura, J. D.; Luty, B. A.; McCammon, J. A. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comp. Phys. Commun.* **1991**, *62*, 187-197.

Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions. I: The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425-445.

Evers, A.; Gohlke, H.; Klebe, G. Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. *J. Mol. Biol.* **2003**, *334*, 327-345.

Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175-1189.

Finn, P. W.; Kavraki, L. E. Computational approaches to drug design. *Algorithmica* **1999**, *25*, 347-371.

Fiser, A.; Do, R. K.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9*, 1753-1773.

Flohil, J. A.; Vriend, G.; Berendsen, H. J. C. Completion and refinement of 3-D homology models with restricted molecular dynamics: Application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Prot. Struct. Func. Gen.* **2002**, *48*, 593-604.

Folkman, J.; D'Amore, P. A. Blood vessel formation: what is its molecular basis? *Cell* **1996**, *87*, 1153-1155.

Forster, M. J. Molecular modelling in structural biology. *Micron* **2002**, *33*, 365-384.

Frenkel, D.; Smit, B. *Understanding molecular simulation. From algorithms to applications*, 2nd ed.; Academic Press: London, 2002.

Gabius, H. J. Tumor lectinology: at the intersection of carbohydrate chemistry, biochemistry, cell biology and oncology. *Angew. Chem. Int. Ed. Eng.* **1988**, *27*, 1267-1276.

Gelhaar, D. K.; Bouzida, D.; Rejto, P. A. In *Rational Drug Design: Novel methodology and practical applications*; Parrill, L., Reddy, M. R., Eds.; Americal Chemical Society: Washington, DC, 1999; pp 292-311.

Gelhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 Protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317-324.

Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: recent developments in the *de novo* design of molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207-217.

Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: a program for structure generation. *J. Comput. Aided Mol. Des* **1993**, *7*, 127-153.

Glover, F.; Laguna, M. Tabu search. In *Modern Heuristic Techniques for Combinatorial Problems*; Reeves, C. R., Ed.; Blackwell Scientific Publications: Oxford, 1993; 70-150.

Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Prot. Struct. Func. Gen.* **1994**, *18*, 309-317.

Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337-356.

Goodford, P. J. A computational procedure for determining energetically favorable binding-sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.

Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Prot. Struct. Func. Gen.* **1990**, *8*, 195-202.

Gorre, M. E.; Mohammed, M.; Ellwood, K.; Hsu, N.; Paquette, R.; Rao, P. N.; Sawyers, C. L. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **2001**, *293*, 876-880.

Gou, Z. Y.; Thirumalai, D. Kinetics of Protein Folding: Nucleation mechanism, time scales and pathways. *Biopolymers* **1995**, *36*, 83-102.

Gould, C.; Wong, C. F. Designing specific protein kinase inhibitors: Insights from computer simulations and comparative sequence/structure analysis. *Pharmacol. Ther.* **2002**, *93*, 169-178.

Graves, B. J.; Crowther, R. L.; Chandran, C.; Rumberger, J. M.; Li, S.; Huang, K.-S.; Presky, D. H.; Familletti, P. C.; Wolitzky, B. A.; Burns, D. K. Insight into E-selectin/ligand interaction from the crystal structure and mutagenesis of the lec/EGF domains. *Nature* **1994**, *367*, 532-538.

Gray, N. S.; Wodicka, L.; Thunnissen, A. M.; Norman, T. C.; Kwon, S.; Espinoza, F. H.; Morgan, D. O.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S. H.; Lockhart, D. J.; Schultz, P. G. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **1998**, *281*, 533-538.

Guex, N.; Diemand, A.; Peitsch, M. C. Protein modelling for all. *TiBS* **1999**, *24*, 364-367.

Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714-2723.

Halgren, T. A. Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490-519.

Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Prot. Struct. Func. Gen.* **2002**, *47*, 409-443.

Hanahan, D.; Folkman, J. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell* **1996**, *86*, 353-364.

Harpur, A. G.; Andres, A. C.; Ziemiecki, A.; Aston, R. R.; Wilks, A. F. Jak2, a 3rd member of the Jak family of protein tyrosine kinases. *Oncogene* **1992**, *7*, 1347-1353.

Hart, T. N.; Read, R. J. A multiple-start Monte-Carlo docking method. *Prot. Struct. Func. Gen.* **1992**, *13*, 206-222.

Heldin, C. H. Dimerization of cell-surface receptors in signal-transduction. *Cell* **1995**, *80*, 213-223.

Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE**:** Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359-363.

Holm, L.; Sander, C. Protein-structure comparison by alignment of distance matrices. *J. Mol. Biol.* **1993**, *233*, 123-138.

Holm, L.; Sander, C. New structure: Novel fold? *Structure* **1997**, *5*, 165-171.

Holm, L.; Sander, C. Dictionary of recurrent domains in protein structures. *Prot. Struct. Func. Gen.* **1998**, *33*, 88-96. (a)

Holm, L.; Sander, C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **1998**, *26*, 316-319. (b)

Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144-1149.

Hooft, R. W. W.; Sander, C.; Vriend, G. Verification of protein structures: side-chain planarity. *J. Appl. Crystallogr.* **1996**, *29*, 714-716.

Hornak, V.; Simmerling, C. Generation of accurate protein loop conformations through low-barrier molecular dynamics. *Prot. Struct. Func. Gen.* **2003**, *51*, 577-590.

Høy, M. Building simple, reliable and relevant multivariate data-analysis tools, Ph.D. Thesis, Department of Chemistry, Norwegian University of Science and Technology, 2002.

Hubbard, S. R. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.* **1997**, *16*, 5573-5581.

Hubbard, S. R.; Till, J. H. Protein tyrosine kinase structure and function. *Annu. Rev. Biochem.* **2000**, *69*, 373-398.

Hubbard, S. R.; Wei, L.; Elis, L.; Hendrickson, W. A. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* **1994**, *372*, 746-754.

Huber, T.; Torda, A. E.; van Gunsteren, W. F. Structure optimization combining soft-core interaction functions, the diffusion equation method, and molecular dynamics. *J. Phys. Chem. A* **1997**, *101*, 5926-5930.

Huwe, C. M.; Woltering, T. J.; Jiricek, J.; Weitz-Schmidt, G.; Wong, C. H. Design, synthesis and biological evaluation of aryl-substituted sialyl Lewis x mimetics prepared via cross-metathesis of *C*-fucopeptides. *Bioorgan. Med. Chem.* **1999**, *7*, 773-788.

Ihle, J. N.; Witthuhn, B. A.; Quelle, F. W.; Yamamoto, K.; Silvennoinen, O. Signaling through the hematopoietic cytokine receptors. *Annu. Rev. Immunol.* **1995**, *13*, 369-398.

Ishima, R.; Torchia, D. A. Protein dynamics from NMR. *Nat. Struct. Biol.* **2000**, *7*, 740-743.

Jensen, F. *Introduction to computational chemistry*; John Wiley & Sons, Ltd.: Chichester, 1999.

Ji, H. T.; Zhang, W. N.; Zhang, M.; Kudo, M.; Aoyama, Y.; Yoshida, Y.; Sheng, C. Q.; Song, Y. L.; Yang, S.; Zhou, Y. J.; Lu, J. G.; Zhu, J. Structure-based *de novo* design, synthesis, and biological evaluation of non-azole inhibitors specific for lanosterol 14 alpha-demethylase of fungi. *J. Med. Chem.* **2003**, *46*, 474-485.

Jiang, F.; Kim, S.-H. "Soft docking": Matching of molecular-surface cubes. *J. Mol. Biol.* **1991**, *219*, 79-102.

Jirousek, M. R.; Gillig, J. R.; Gonzalez, C. M.; Heath, W. F.; McDonald, J. H.; Neel, D. A.; Rito, C. J.; Singh, U.; Stramm, L. E.; Melikian-Badalian, A.; Baevsky, M.; Ballas, L. M.; Hall, S. E.; Winneroski, L. L.; Faul, M. M. (S)-13-[(dimethylamino)methyl]-10,11,14,15-tetrahydro-4,9:16,21-dimetheno-1H,13H-dibenzo[e,k]pyrrolo[3,4-h][1,4,13]oxadiazacyclohexadecene-1,3(2H)-dione (LY333531) and related analogues: Isozyme selective inhibitors of protein kinase C beta. *J. Med. Chem.* **1996**, *39*, 2664-2671.

Johnson, L. N.; Noble, M. E. M.; Owen, D. J. Active and inactive protein kinases: Structural basis for regulation. *Cell* **1996**, *85*, 149-158.

Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*, 4th ed.; Prentice-Hall, Inc.: New Jersey, 1998.

Kairys, V.; Gilson, M. K. Enhanced docking with the mining minima optimizer: Acceleration and side-chain flexibility. *J. Comput. Chem.* **2002**, *23*, 1656-1670.

Kirken, R. A.; Erwin, R. A.; Taub, D.; Murphy, W. J.; Behbod, F.; Wang, L. H.; Pericle, F.; Farrar, W. L. Tyrphostin AG490 inhibits cytokine-mediated JAK3/STAT5a/b signal transduction and cellular proliferation of antigen-activated human T cells. *J. Leukocyte Biol.* **1999**, *65*, 891-899.

Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimisation by simulated annealing. *Science* **1983**, *220*, 671-680.

Klagsbrun, M.; D'Amore, P. A. Regulators of angiogenesis. *Annu. Rev. Physiol.* **1991**, *53*, 217-239.

Klagsbrun, M.; Edelman, E. R. Biological and biochemical properties of fibroblast growth factors. Implications for the pathogenesis of atherosclerosis. *Arteriosclerosis* **1989**, *9*, 269-278.

Klebe, G. Recent developments in structure-based drug design. *J. Mol. Med.-JMM* **2000**, *78*, 269-281.

Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indexes in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130-4146.

Klebe, G.; Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput. Aided Mol. Des.* **1994,** *8*, 583-606.

Klopocki, A. G.; Laskowska, A.; Antoniewicz-Papis, J.; Duk, M.; Lisowska, E.; Ugorski, M. Role of sialosyl Lewis^a in adhesion of colon cancer cells. The antisense RNA approach. *Eur. J. Biochem.* **1998**, *253*, 309-318.

Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424-440.

Koehl, P.; Levitt, M. A brighter future for protein structure prediction. *Nat. Struct. Biol.* **1999**, *6*, 108-111.

Kolodny, R.; Koehl, P.; Guibas, L.; Levitt, M. Small libraries of protein fragments model native protein structures accurately. *J. Mol. Biol.* **2002**, *323*, 297-307.

Kuiper, R. A. J.; Schellens, J. H. M.; Blijham, G. H.; Beijnen, J. H.; Voest, E. E. Clinical research on antiangiogenic therapy. *Pharmacol. Res.* **1998**, *37*, 1-16.

Kumar, R.; Fidler, I. J. Angiogenic molecules and cancer metastasis. *In Vivo* **1998**, *18*, 27-34.

Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283-291.

Laskowski, R. A.; Rullmann, J. A. C.; MacArthur, M. W.; Kaptein, R. ; Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **1996**, *8*, 477-486.

le Grand, S. M.; Merz Jr., K. M. Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. *J. Comput. Chem.* **1993**, *14*, 349-352.

Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345-356.

Leach, A. R. *Molecular modelling. Principles and applications*, 2nd ed.; Prentice Hall: Essex, 2001.

Lesk, A. M.; Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **1980**, *136*, 225-270.

Levitt, M. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **1992**, *226*, 507-533.

Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884-1897.

Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc.* **2002**, *124*, 5632-5633.

Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers* **2003**, *68*, 47-62.

Lindauer, K.; Loerting, T.; Liedl, K. R.; Kroemer, R. T. Prediction of the structure of human Janus kinase 2 (JAK2) comprising the two carboxy-terminal domains reveals a mechanism for autoregulation. *Protein Eng.* **2001**, *14*, 27-37.

Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **1997**, *23*, 3-25.

Liu, H.; Duan, Z.; Luo, Q.; Shi, Y. Structure-based ligand design by dynamically assembling molecular building blocks at binding site. *Proteins* **1999**, *36*, 462-470.

Liu, H.-Y.; Kuntz, I. D.; Zou, X. Pairwise GB/SA scoring function for structure-based drug design. *J. Phys. Chem. B* **2004**, *108*, 5453-5462.

Liu, J.; Tøndel, K.; Adcock, S.; Gribskov, M.; Niedner, H. R.; McCammon, J. A.; Nielsen, J. E. Homology modelling of protein kinases. *In Preparation.*

LoConte, L.; Ailey, B.; Hubbard, T. J. P.; Brenner, S. E.; Murzin, A. G.; Chothia, C. SCOP**:** A structural classification of proteins database. *Nucleic Acids Res.* **2000**, *28*, 257-259.

Lüthy, R.; Bowie, J. U.; Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **1992**, *356*, 83-85.

Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, *7*, 1047-1055.

Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Sci.* **2002**, *11*, 184-197.

MacKerell Jr., A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr., R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102***,** 3586-3616. (a)

MacKerell, A. D.; Brooks, B.; Brook III, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMm: The energy function and its parameterization. In *Encyclopedia of Computational Chemistry*, Schleyer, P. R., Ed.; John Wiley & Sons: New York, 1998; pp. 271-277. (b)

Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.* **1995**, *91*, 57-95.

Makino, S.; Kuntz, I. D. Automated flexible ligand docking method and its application for database search. *J. Comput. Chem.* **1997**, *18*, 1812-1825.

Mancera, R. L.; Källblad, P.; Todorov, N. P. Ligand-protein docking using a quantum stochastic tunneling optimization method. *J. Comput. Chem.* **2004**, *25*, 858-864.

Manne, R. Analysis of two Partial-Least-Squares algorithms for multivariate calibration. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 187-197.

Martens, C. L.; Cwirla, S. E.; Lee, R. Y. W.; Whitehorn, E.; Chen, E. Y. F.; Bakker, A.; Martin, E. L.; Wagstrom, C.; Gopalan, P.; Smith, C. W.; Tate, E.; Koller, K. J.; Schatz, P. J.; Dower, W. J.; Barrett, R. W. Peptides which bind to E-selectin and block neutrophil adhesion. *J. Biol. Chem.* **1995**, *270*, 21129-21136.

Martens, H.; Martens, M. *Multivariate analysis of quality. An introduction*; John Wiley & Sons, Ltd.: Chichester, 2000; pp 111-125.

Marti-Renom, M. A.; Stuart, A.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291-325.

Massova, I.; Kollman, P. A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* **1999**, *121*, 8133-8143.

McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76-90.

McPherson, J. D.; Marra, M.; Hillier, L.; Waterston, R. H.; Chinwalla, A.; Wallis, J.; Sekhon, M.; Wylie, K.; Mardis, E. R.; Wilson, R. K.; Fulton, R.; Kucaba, T. A.; Wagner-McPherson, C.; Barbazuk, W. B.; Gregory, S. G.; Humphray, S. J.; French, L.; Evans, R. S.; Bethel, G.; Whittaker, A.; Holden, J. L.; McCann, O. T.; Dunham, A.; Soderlund, C.; Scott, C. E.; Bentley, D. R.; Schuler, G.; Chen, H.-C.; Jang, W. H.; Green, E. D.; Idol, J. R.; Maduro, V. V. B.; Montgomery, K. T.; Lee, E.; Miller, A.; Emerling, S.; Kucherlapati, R.; Gibbs, R.; Scherer, S.; Gorrell, J. H.; Sodergren, E.; Clerc-Blankenburg, K.; Tabor, P.; Naylor, S.; Garcia, D.; de Jong, P. J.; Catanese, J. J.; Nowak, N.; Osoegawa, K.; Qin, S. Z.; Rowen, L.; Madan, A.; Dors, M.; Hood, L.; Trask, B.; Friedman, C.; Massa, H.; Cheung, V. G.; Kirsch, I. R.; Reid, T.; Yonescu, R.; Weissenbach, J.; Bruls, T.; Heilig, R.; Branscomb, E.; Olsen, A.; Doggett, N.; Cheng, J.-F.; Hawkins, T.; Myers, R. M.; Shang, J.; Ramirez, L.; Schmutz, J.; Velasquez, O.; Dixon, K.; Stone, N. E.; Cox, D. R.; Haussler, D.; Kent, W. J.; Furey, T.; Rogic, S.; Kennedy, S.; Jones, S.; Rosenthal, A.; Wen, G. P.; Schilhabel, M.; Gloeckner, G.; Nyakatura, G.; Siebert, R.; Schlegelberger, B.; Korenberg, J.; Chen, X.-N.; Fujiyama, A.; Hattori, M.; Toyoda, A.; Yada, T.; Park, H.-S.; Sakaki, Y.; Shimizu, N.; Asakawa, S.; Kawasaki, K.; Sasaki, T.; Shintani, A.; Shimizu, A.; Shibuya, K.; Kudoh, J.; Minoshima, S.; Ramser, J.; Seranski, P.; Hoff, C.; Poustka, A.; Reinhardt, R.; Lehrach, H. A physical map of the human genome. *Nature* **2001**, *409*, 934-941.

Meirovitch, H. Calculation of the free energy and the entropy of macromolecular systems by computer simulation. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1998; 11, pp 1-74.

Melo, F.; Feytmans, E. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **1998**, *277*, 1141-1152.

Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505-524.

Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

Meydan, N.; Grunberger, T.; Dadi, H.; Shahar, M.; Arpaia, E.; Lapidot, Z.; Leeder, J. S.; Freedman, M.; Cohen, A.; Gazit, A.; Levitzki, A.; Roifman, C. M. Inhibition of acute lymphoblastic leukaemia by a Jak-2 inhibitor. *Nature* **1996**, *379*, 645-648.

Miller, M. A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* **2002**, *1,* 220-227.

Miranker, A.; Karplus, M. Functionality maps of binding-sites - a multiple copy simultaneous search method. *Prot. Struct. Func. Gen.* **1991**, *11*, 29-34.

Mohammadi, M.; Froum, S.; Hamby, J. M.; Schroeder, M. C.; Panek, R. L.; Lu, G. H.; Eliseenkova, A. V.; Green, D.; Schlessinger, J.; Hubbard, S. R. Crystal structure of an angiogenesis inhibitor bound to the FGF receptor tyrosine kinase domain. EMBO J. **1998**, *17*, 5896-5904.

Molecular Operating Environment™, Version 2002.03, *Chemical Computing Group, Inc.*, 2002.

Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.

Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* **1996,** *10*, 293-304.

Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP): Round IV. *Prot. Struct. Func. Gen.* **2001**, *Suppl. 5*, 2-7.

Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Prot. Struct. Func. Gen.* **2003**, *53, Suppl. 6*, 334-339.

Muegge, I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspect. Drug Discov.* **2000**, *20*, 99-114.

Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418-425.

Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791-804.

Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP **-** A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536-540.

Ng, K. K.-S.; Weis, W. I. Structure of a selectin-like mutant of mannose-binding protein complexed with sialylated and sulfated Lewis^x oligosaccharides. *Biochemistry* **1997**, *36*, 979-987.

Nicholls, A.; Sharp, K. A.; Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Prot. Struct. Func. Gen.* **1991**, *11*, 281-296.

Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity - A review. QSAR Comb. Sci. **2003**, *22*, 1006-1026.

Nissink, J. W. M.; Verdonk, M. L.; Klebe, G. Simple knowledge-based descriptors to predict protein-ligand interactions. Methodology and validation. *J. Comput. Aided Mol. Des.* **2000**, *14*, 787-803.

Oldfield, T. J. Squid: A program for the analysis and display of data from crystallography and molecular dynamics. *J. Mol. Graph.* **1992**, *10*, 247-252.

Österberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Prot. Struct. Func. Gen.* **2002**, *46*, 34-40.

Pastor, M.; Cruciani, G. A novel strategy for improving ligand selectivity in receptor-based drug design. *J. Med. Chem.* **1995**, *38*, 4637-4647.

Pautsch, A.; Zoephel, A.; Ahorn, H.; Spevak, W.; Hauptmann, R.; Nar, H. Crystal structure of bisphosphorylated IGF-1 receptor kinase: Insight into domain movements upon kinase activation. *Structure* **2001**, *9*, 955-965.

Pearl, F. M. G.; Lee, D.; Bray, J. E.; Sillitoe, I.; Todd, A. E.; Harrison, A. P.; Thornton, J. M.; Orengo, C. A. Assigning genomic sequences to CATH. *Nucleic Acids Res.* **2000**, *28*, 277-282.

Pearlman, D. A.; Charifson, P. A. Improved scoring of ligand-protein interactions using OWFEG free energy grids. *J. Med. Chem.* **2001**, *44*, 502-511.

Pegg, S. C. H.; Haresco, J. J.; Kuntz, I. D. A genetic algorithm for structure-based *de novo* design. *J. Comput. Aided Mol. Des.* **2001**, *15*, 911-933.

Peitsch, M. C. Protein modeling by E-mail. *Bio/Technology* **1995**, *13*, 658-660.

Peitsch, M. C. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc.Trans*. **1996**, *24*, 274-279.

Pellegrini, S.; Dusanter-Fourt, I. The structure, regulation and function of the Janus kinases (JAKs) and the signal transducers and activators of transcription (STATs). *Eur. J. Biochem.* **1997**, *48*, 615-633.

Pepper, M. S. Positive and negative regulation of angiogenesis: from cell biology to the clinic. *Vasc. Med*. **1996**, *1*, 259-266.

Peters, K. P.; Fauck, J.; Frommel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **1996**, *256*, 201-213.

Qian, B.; Goldstein, R. A. Optimization of a new score function for the generation of accurate alignments. *Prot. Struct. Func. Gen.* **2002**, *48*, 605-610.

Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997**, *101*, 3005-3014.

Rappé, A. K.; Casewit, C. J. *Molecular mechanics across chemistry*; University Science Books: Sausalito, California, 1997.

Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: Placing discrete water molecules during protein-ligand docking predictions. *Prot. Struct. Func. Gen.* **1999**, *34*, 17-28.

Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.

Rarey, M.; Lengauer, T. A recursive algorithm for efficient combinatorial library docking. *Perspect. Drug Discov.* **2000**, *20*, 63-81.

Raz, A.; Zhu, D. G.; Hogan, V.; Shah, N.; Raz, T.; Karkash, R.; Pazerini, G.; Carmi, P. Evidence for the role of 34-KDa galactoside-binding lectin in transformation and metastasis. *Int. J. Cancer* **1990**, *46*, 871-877.

Read, R. J.; Brayer, G. D.; Jurásek, L.; James, M. N. G. Critical evaluation of comparative model building of Streptomyces griseus trypsin. *Biochemistry* **1984**, *23*, 6570-6575.

Revelle, B. M.; Scott, D.; Beck, P. J. Single amino acid residues in the E- and P-selectin epidermal growth factor domains can determine carbohydrate binding specificity. *J. Biol. Chem.* **1996**, *271*, 16160-16170.

Richmond, T. J. Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.* **1984**, *178*, 63-88.

Richter, M. F.; Dumenil, G.; Uze, G.; Fellous, M.; Pellegrini, S. Specific contribution of Tyk2 JH regions to the binding and the expression of the interferon alpha/beta receptor component IFNAR1. *J. Biol. Chem.* **1998**, *273*, 24723-24729.

Ripka, A. S.; Satyshur, K. A.; Bohacek, R. S.; Rich, D. H. Aspartic protease inhibitors designed from computer-generated templates bind as predicted. *Org. Lett.* **2001**, *3*, 2309-2312.

Robinson, M. K.; Stephens, P. E. Neutrophil adhesion: a point for therapeutic intervention? *Curr. Opin. Biotechnol.* **1992**, *3*, 662-667.

Russell, R. B.; Barton, G. J. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* **1994**, *244*, 332-350.

Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779-815.

Sánchez, R.; Sali, A. Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.* **1997**, *7*, 206-214.

Sawyer, T. K. Drug design - Chemistry and biology. *Biotechniques* **2001**, *31*, 1164-1171.
Scarsi, M.; Majeux, N.; Caflisch, A. Hydrophobicity at the surface of proteins. *Prot. Struct. Func. Gen.* **1999**, *37*, 565-575.

Schaffer, L.; Verkhivker, G. M. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Prot. Struct. Func. Gen.* **1998**, *33*, 295-310.

Schafferhans, A.; Klebe, G. Docking ligands onto binding site representations derived from proteins built by homology modelling. *J. Mol. Biol.* **2001**, *307*, 407-427.

Scheffler, K.; Ernst, B.; Katopodis, A.; Magnani, J. L.; Wang, W. T.; Weisemann, R.; Peters, T. Determination of the bioactive conformation of the carbohydrate ligand in the E-selectin/sialyl Lewis^x complex. *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 1841-1844.

Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Structural mechanism for STI-571 inhibition of Abelson tyrosine kinase. *Science* **2000**, *289*, 1938-1942.

Schlessinger, J.; Ullrich, A. Growth-factor signaling by receptor tyrosine kinases. *Neuron* **1992**, *9*, 383-391.

Schnecke, V.; Kuhn, L. A. Database screening for HIV protease ligands: the influence of binding-site conformation and representation on ligand selectivity. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1999**, 242-251. (a)

Schnecke, V.; Kuhn, L. A. In *Rigidity Theory and Applications*; Thorpe, M. F., Duxbury, P. M., Eds.; Kluwer Academic/Plenum Publishers: New York, 1999; pp 385-400. (b)

Schnecke, V.; Kuhn, L. A. Virtual screening with solvation and ligand-induced complementarity. *Perspect. Drug Discov.* **2000**, *20*, 171-190.

Schnecke, V.; Swanson, C. A.; Getzoff, E. D.; Tainer, J. A.; Kuhn, L. A. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Prot. Struct. Func. Gen.* **1998**, *33*, 74-87.

Schneider, G.; Böhm, H.-J. Virtual screening and fast automated docking methods. *Drug Discov. Today* **2002**, *7*, 64-70.

Schonbrun, J.; Wedemeyer, W. J.; Baker, D. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **2002**, *12*, 348-354.

Schwartz, R. M.; Dayhoff, M. O. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. A perspective is derived from protein and nucleic acid sequence data. *Science* **1978**, *199*, 395-403.

Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **2003,** *31*, 3381-3385.

Shoichet, B. K.; McGovern, S. L.; Wei, B. Q.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439-446.

Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Prot. Struct. Func. Gen.* **1993**, *17*, 355-362.

Smith, A. Navigating the evolving world of drug discovery. *Nat. Rev. Drug Discov.* **2002**, *1*, 3.

Somers, W. S.; Tang, J.; Shaw, G. D.; Camphausen, R. T. Insights into the molecular basis of leukocyte tethering and rolling revealed by structures of P-and E-Selectin bound to SLe(X) and PSGL-1. *Cell* **2000,** *103*, 467-479.

Sotriffer, C.; Klebe, G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco* **2002**, *57*, 243-251.

Stahn, R.; Schafer, H.; Kernchen, F.; Schreiber, J. Multivalent sialyl Lewis x ligands of definite structures as inhibitors of E-selectin mediated cell adhesion. *Glycobiology* **1998**, *8*, 311-319.

Stahura, F. L.; Bajorath, J. Bio- and chemo-informatics beyond data management: crucial challenges and future opportunities. *Drug Discov. Today* **2002**, *7, Suppl. S.*, S41-S47.

Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127-6129.

Stultz, C. M.; Karplus, M. MCSS functionality maps for a flexible protein. *Prot. Struct. Func. Gen.* **1999**, *37*, 512-529.

Swiss-PdbViewer, Version 3.7b2, *Glaxo Wellcome Experimental Research*, 2001.

Szekanecz, Z.; Szegedi, G.; Koch, A. E. Angiogenesis in rheumatoid arthritis: pathogenic and clinical significance. *J. Investig. Med.* **1998**, *46*, 27-41.

Tappura, K.; Lahtela-Kakkonen, M.; Teleman, O. A new soft-core potential function for molecular dynamics applied to the prediction of protein loop conformations. *J. Comput. Chem.* **2000**, *21*, 388-397.

Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des* **2002**, *16*, 151-166.

Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. FDS: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J. Comput. Chem.* **2003**, *24*, 1637-1656.

Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 527-541.

Teodoro, M. L.; Kavraki, L. E. Conformational flexibility models for the receptor in structure based drug design. *Curr. Pharm. Design* **2003**, *9*, 1635-1648.

Terfloth, L.; Gasteiger, J. Neural networks and genetic algorithms in drug design. *Drug Discov. Today* **2001**, *6*, *Suppl. S*, S102-S108.

Todorov, N. P.; Mancera, R. L.; Monthoux, P. H. A new quantum stochastic tunnelling optimisation method for protein-ligand docking. *Chem. Phys. Lett.* **2003**, *369*, 257-263.

Tolentino, M. J.; Adamis, A. P. Angiogenic factors in the development of diabetic iris neovascularization and retinopathy. *Int. Ophthalmol. Clin.* **1998**, *38*, 77-94.

Topham, C. M.; Srinivasan, N.; Thorpe, C. J.; Overington, J. P.; Kalsheker, N. A. Comparative modelling of major house dust mite allergen der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng.* **1994**, *7*, 869-894.

Torda, A. E. Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **1997**, *7*, 200-205.

Tress, M. L.; Jones, D.; Valencia, A. Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.* **2003**, *330*, 705-718.

van der Geer, P.; Hunter, T.; Lindberg, R. A. Receptor protein-tyrosine kinases and their signal-transduction pathways. *Annu. Rev. Cell Biol.* **1994**, *10*, 251-337.

Venclovas, C.; Zemla, A.; Fidelis, K.; Moult, J. Criteria for evaluating protein structures derived from comparative modeling. *Prot. Struct. Func. Gen.* **1997**, *Suppl. 1*, 7-13.

Verdonk, M. L.; Cole, J. C.; Taylor, R. SUPERSTAR: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **1999**, *289*, 1093-1108.

Verdonk, M. L.; Cole, J. C.; Watson, P.; Gillet, V.; Willett, P. SUPERSTAR: improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol.* **2001**, *307*, 841-859.

Vodovozova, E. L.; Moiseeva, E. V.; Grechko, G. K.; Gayenko, G. P.; Nifant'ev, N. E.; Bovin, N. V.; Molotkovsky, J. G. Antitumour activity of cytotoxic liposomes equipped with selectin ligand SiaLe^x, in a mouse mammary adenocarcinoma model. *Eur. J. Cancer* **2000**, *36*, 942-949.

Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8*, 52-56.

Wang, L. H.; Kirken, R. A.; Erwin, R. A.; Yu, C. R.; Farrar, W. L. JAK3, STAT, and MAPK signaling pathways as novel molecular targets for the tyrphostin AG490 regulation of IL-2-mediated T cell response. *J. Immunol.* **1999**, *162*, 3897-3904.

Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11-26.

Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.* **1998**, *4*, 379-394.

Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287-2303.

Wang, R; Gao, Y.; Lai, L. LigBuilder: A multi-purpose program for structure-based drug design. *J. Mol. Model.* **2000**, *6*, 498-516.

Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Amer. Chem. Soc.* **1984**, *106,* 765-784.

Wenzel, W.; Hamacher, K. Stochastic tunneling approach for global minimization of complex potential energy landscapes. *Phys. Rev. Lett.* **1999**, *82*, 3003-3007.

Wojciechowski, M.; Skolnick, J. Docking of small ligands to low-resolution and theoretically predicted receptor structures. *J. Comput. Chem.* **2002**, *23*, 189-197.

Wong, C. F.; Hünenberger, P. H.; Akamine, P.; Narayana, N.; Diller, T.; McCammon, J. A.; Taylor, S.; Xuong, N. H. Computational analysis of PKA-balanol interactions. *J. Med. Chem.* **2001**, *44*, 1530-1539.

Wong, C. F.; McCammon, J. A. Protein flexibility and computer-aided drug design. *Annu. Rev. Pharmacol. Toxicol.* **2003**, *43*, 31-45.

Wong, C. F.; Thacher, T.; Rabitz, H. Sensitivity analysis in biomolecular simulation. In *Reviews in Computational Chemistry*, Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1998; 12, pp 281- 326.

Xuan, Y. T.; Guo, Y. R.; Han, H.; Zhu, Y. Q.; Bolli, R. An essential role of the JAK-STAT pathway in ischemic preconditioning. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 9050-9055.

Yan, H.; Piazza, F.; Krishnan, K.; Pine, R.; Krolewski, J. J. Definition of the interferon-α receptor-binding domain on the TYK2 kinase. *J. Biol. Chem.* **1998**, *273*, 4046-4051.

Yoon, S.; Welsh, W. J. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 88-96.

Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 883-902.

Zhu, J.; Fan, H.; Liu, H.; Shi, Y. Structure-based ligand design for flexible proteins: Application of new F-DycoBlock. *J. Comput. Aided Mol. Des.* **2001**, *15*, 979-996.

Zimmermann, J.; Buchdunger, E.; Mett, H.; Meyer, T.; Lydon, N. B. Potent and selective inhibitors of the Abl-kinase: Phenylamino-pyrimidine (PAP) derivatives. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 187-192.

Zou, X. Q.; Sun, Y. X.; Kuntz, I. D. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J. Am. Chem. Soc.* **1999**, *121*, 8033-8043.

Paper I

Prediction of homology model quality with multivariate regression

Kristin Tøndel

Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, Norway
E-mail: kristito@phys.chem.ntnu.no
Phone: +47 73 59 41 73
Fax: +47 73 59 16 76

Abstract

A new method has been developed for prediction of homology model quality directly from the sequence alignment, using multivariate regression. Hence, the expected quality of future homology models can be estimated using only information about the primary structure. This method has been applied to protein kinases, and can easily be extended to other protein families. Prior to the multivariate regression analysis, a set of homology models was verified by comparison to experimental structures. The homology model quality was evaluated by calculation of root mean square deviations (RMSDs) and comparison of inter-residue contact areas. The homology model quality measures were used as dependent variables in a Partial Least Squares (PLS) regression, using a matrix of alignment score profiles found from the Point Accepted Mutation (PAM) 250 similarity matrix as independent variables. The method presented here can be used to effectively choose the correct templates to use for the homology modelling, and to identify regions of the protein structure that are difficult to model, as well as alignment errors. Hence, this method is a useful tool for assuring that the best possible homology model is generated.

Key Words: Homology modelling, homology model quality prediction, inter-residue contact areas, modelling templates, multivariate regression

1 Introduction

During the last decade, homology modelling of protein structures has become a commonly used technique. Homology modelling is the procedure of generating a model of a protein using an experimental structure of a related protein as a template.[1, 2, 3, 4] Many different programs are available for this purpose, and homology models of proteins are currently used in a wide variety of disciplines, ranging from drug design, to studies of mutations and protein engineering.[1] With the user-friendly modelling programs available, constructing a homology model of a protein is straightforward, but the quality of the results may vary a lot since automatic methods not always find optimal alignments or loop predictions, especially when the sequence identity is below 40%.[1, 5] An inaccurate homology model may be misleading, because relatively small structural errors may lead to large errors in e.g. binding energy calculations. Accuracies of the various homology model building methods are relatively similar when used optimally.[1, 6] Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy. Even homology models generated from very high quality sequence alignments might contain severe errors.[7] Hence, it is important to evaluate the quality of homology models made from high quality sequence alignments, and to be able to predict the model quality for a given target-template pair. This is important both in order to select the correct template structures to use for the homology modelling, and to evaluate whether useful information can be extracted from a future homology model. In this way, we can avoid spending time on generation of low-quality homology models.

Models of three-dimensional (3D) protein structures can be evaluated according to a variety of criteria, such as stereochemistry (bond lengths, bond angles, torsion angles etc.), packing, formation of a hydrophobic core, residue and atomic solvent accessibilities, spatial distribution of charged groups, distribution of atom-atom distances, atomic volumes and main-chain hydrogen bonding.[8, 9, 10, 11] Large deviations from the most likely values have been interpreted as indicators of errors in the model structure. Methods based on 3D profiles and statistical potentials of mean force also exist, that take many of these criteria into account implicitly.[12, 13, 14, 15] These methods evaluate the environment of each residue as seen in the model, compared to the expected environment as observed in experimental structures.

The accuracy of protein structure models can also be evaluated by comparison to experimental structures of the targets.[16, 17, 18, 19] The most common method for comparison of two 3D structures is calculation of root mean square deviations (RMSDs) between corresponding atoms in the structures. However, the geometric measures only provide meaningful results when the entire extent of the proteins is comparable. For example, a set of partially correct structures cannot be ranked because the incorrect portions will dominate an RMSD value. When restricting the comparison to certain parts of the structure, the choice of relevant parts may also be somewhat arbitrary. An alternative is to use the surface area of residue contacts, which does not require a superpositioning of the structures that are being compared. A new surface area based comparison method has been developed.[7] This method is similar to the Contact Area Difference (CAD) number,[20] but differs in both technical details and in the definition of a single scalar value to quantify the similarity. The surface areas are calculated using a Boolean logic based algorithm.[21] A two-dimensional matrix is constructed by calculating every pairwise contact surface area between the residues in each protein structure. When two protein structures are compared, the difference between the contact area matrices for the two structures is calculated. The elements in the resulting matrix are negative for incorrectly occurring and overestimated contacts, zero for correct contacts and non-contacting residue pairs, and positive for underestimated or missing contacts in the model structure. In the following, this matrix will be referred to as the inter-residue contact area error matrix. Analysis of residue-residue contacts has been used to evaluate structure predictions,[22] and the conservation of side-chain interactions in homologous proteins.[23] Contact-based measures can also be applied to simplified protein descriptions.[24]

All methods mentioned above for evaluation of the quality of protein structure models operate on the 3D models themselves. No methods exist that predict the model quality prior to the actual model building. Sequence identity between the target and template above 30% is a relatively good indicator of the expected homology model accuracy, but when the sequence identity is below 30% it becomes unreliable as a measure of the expected model quality.[1] Predicting homology model quality is a difficult task, and can only be done within a specific protein family, since the effective mutation rate, the number and size of insertions and deletions, the number of surface loops, etc. vary between protein families.[25] Even within a specific protein structure, some regions can be modelled with high accuracy, while others are more difficult to model. Loop modelling is known to be a difficult task, and much research is devoted to this part of the homology modelling procedure.[1, 26, 27] Loop modelling techniques range from searching databases of known protein structures for loops having similar end points, to molecular dynamics simulations.[1, 28]

Prediction of the expected homology model quality, given a specific relationship between the primary structure of the target and template proteins, is useful for evaluating whether a homology model can be generated that suits the needs of the specific task. In some cases a model of very high accuracy is needed, while in other cases a model of lower quality can provide sufficient information. Misalignments are the largest source of errors in comparative modelling.[1] In this work, a new method for prediction of homology model accuracy has been developed, that operates only on the target-template sequence alignment. Hence, no information about the 3D structure is needed, and the homology model quality can be predicted for a wide range of sequence identities. This method has been applied to the protein kinase family, but can easily be extended to other protein families. RMSD values between the homology model structures and experimental structures of the same proteins, and differences in inter-residue contact areas between the models and the target X-ray structures are used as measures of the model quality. The method presented here can be used to assure that the correct templates and alignments are chosen, so that the best possible homology model is generated. It is also useful for identification of regions that are difficult to model, as well as errors in the alignment. Possibilities for improving the homology model quality by combination of several homology models are also discussed.

2 Methodology

A regression model has been developed for prediction of the accuracy of homology models of protein structures. This regression model was trained on 292 homology models of proteins for which experimental structures were available for comparison. Calculated RMSDs and differences in inter-residue contact areas between the homology models and the target X-ray structures were used as measures of the accuracy of the homology models. The homology model quality data were used as dependent variables (**Y**) in the Partial Least Squares (PLS) regression analysis.

A matrix of alignment score profiles describing the similarity between the target and template amino acid sequences for each homology model was used as independent variables (**X**) in the regression analysis. Each element in this alignment score matrix contained the value of the Point Accepted Mutation (PAM) 250 similarity matrix[29] for a pair of amino acids that correspond to each other in the sequence alignment. Hence, for each homology model, a score value (corresponding to the PAM250 matrix value) for each pair of residues aligned in the sequence alignment used for the modelling was found, resulting in a matrix of alignment scores, as illustrated in Figure 1. This describes how similar the target and template amino acid sequences are in each position in the alignment.

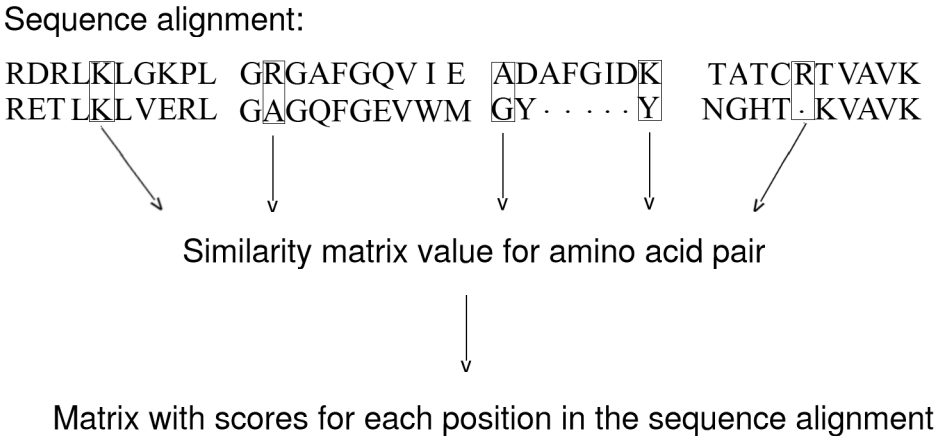


Figure 1. Generation of alignment score profiles.

The PLS regression analysis is illustrated in Figure 2. In this model, the sequence identity between the target and the template and the number of non-modelled residues (caused by gaps in the sequence alignment) are also added to the matrix of independent variables (**X**-matrix). A gap in the sequence alignment appears when an insertion or deletion has occurred during evolution, so that there is a region where the target and template structures differ in length. This often occurs in surface loops where the effective mutation rate is high, meaning that an increased fraction of mutations are leading to functional genes. Gaps in the sequence alignment lead to inaccuracies in the homology modelling, and represent a great challenge when developing homology modelling methods. However, the effective mutation rate in active sites of protein structures is often relatively low,[25] so a homology model might be useful even though a part of the structure (e.g. a surface loop) is not modelled correctly.

Any other available information about the similarity between the target and template structures can also be added to the **X**-matrix to improve the predictive ability of the obtained regression model.

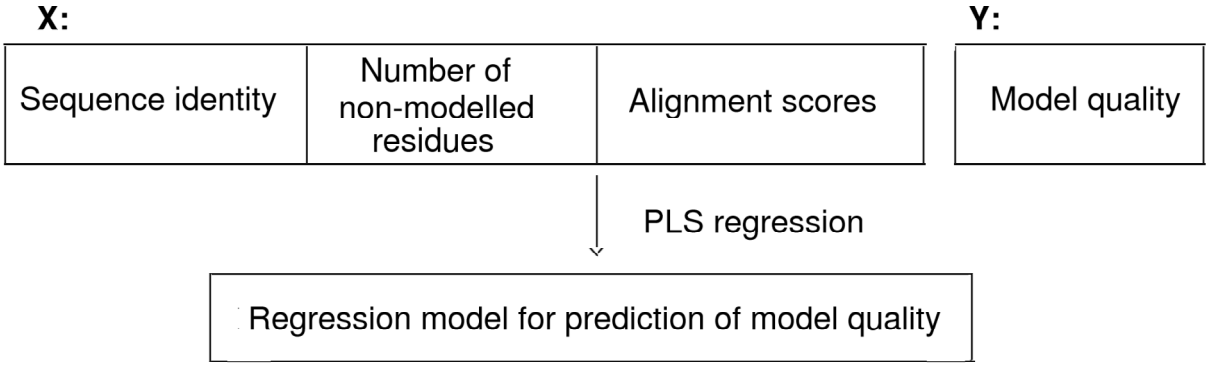


Figure 2. Multivariate analysis of the homology model quality data using alignment scores, sequence identity and number of non-modelled residues as independent variables.

This regression model can be used to predict the accuracy of new homology models. Prior to constructing a new homology model, alignment score profiles can be generated from the sequence alignment between the target and the template. The regression model developed in this work can predict the homology model quality for new homology models from such alignment score profiles. This model can only be used within the protein family for which it has been trained, but similar regression models can be made in the same way for other protein families.

Outliers that should be kept out of the regression analysis can be identified by inspection of influence plots, that is, plots of the residual Y-variance against the leverage. Outliers that have a large effect on the results from the regression analysis will be placed in the upper, right hand part of the influence plot. This can be used to identify members of a protein family that are difficult to model with homology modelling due to large deviations from the other proteins in the family.

The regression coefficients can be used to identify regions that are difficult to model, as well as alignment errors. Regions of the sequence alignment that contain many gaps (regions where the sequence alignment is of low quality) correspond to regions with large variations in the regression coefficients. Comparison of the residuals (for each alignment position) from prediction for a new homology model to the residuals for the homology models included in the regression analysis can also reveal errors in the sequence alignment. Such alignment errors will lead to deviations in the residual pattern.

2.1 Data sets

2.1.1 Protein kinase structures

Two sets of protein kinase structures from the RCSB Protein Data Bank (PDB)[30] were selected for the homology modelling. One set (A) contained fourteen structures with pairwise sequence identities between 14% and 40% (Table 1). This set is a representative set of all protein kinase structures in the PDB. To maximise the structural diversity in the set, all pairs of structures in this set have sequence identity lower than 40%. Only structures with resolution better than 3 Å were selected.

The other set (B) consists of eleven protein kinase structures with pairwise sequence identities of 35-80% (Table 2). This is the sequence identity range where homology modelling is most frequently used. These eleven structures belong to the receptor tyrosine kinase (RTK) family. The RTK family was chosen because it is of great interest in e.g. drug design, and one of the families where experimental structures with the widest range of sequence identities are available in the PDB. The PDB holds 25 entries corresponding to protein kinases in the RTK family. To leave out

multiple X-ray structures of the same proteins, only structures having lower sequence identities than 90% to each other were chosen. This resulted in ten kinase structures, which provide a good representation of the structural diversity in the RTK family. In the interest of exploring the effects of conformational change, the apo-structure of the human insulin receptor protein kinase (HIRPK) was added to the set, so that the set now contained two copies of HIRPK.

For both sets of kinase structures, structures with a ligand in the adenosine triphosphate (ATP)-binding site and high-resolution structures were preferred when multiple structures of the same protein were present in the PDB.

The X-ray structures were superposed using the CE algorithm.[31] Pairwise sequence identities and RMSD values for the two sets of protein kinase structures are given in Table 1 and 2.

Table 1. a) Pairwise sequence identities (%) and b) Cα and Cβ RMSD values (Å) for the fourteen protein kinases in set A*.

a)

PDB entry	1csn__	1b6c_b	1fgi_a	1ir3_a	2src__	1a6o__	1f3m_c	1hck__	1jnk__	1kob_a	1tki_a	1cdk_a	1phk__	1a06__
1csn__	100	21.4	17.6	19.1	19	16.2	18.3	19.7	18.5	17.8	14.1	20.7	17.6	15.9
1b6c_b	21.4	100	30.4	28.8	27	20.4	23	23.6	23.2	21.6	18.3	21.4	23	19.4
1fgi_a	17.6	30.4	100	37	39.8	18.5	23.8	25.5	26.7	22.4	20.1	21.9	19.6	25.4
1ir3_a	19.1	28.8	37	100	40.8	17.7	25	21.9	21.8	20.8	19.7	21.1	20.8	20
2src__	19	27	39.8	40.8	100	18.9	23.4	28.2	22.7	22.4	19.9	22	22.7	21.4
1a6o_	16.2	20.4	18.5	17.7	18.9	100	26.5	32.7	26.5	24	25.3	23.3	26.2	25.4
1f3m_c	18.3	23	23.8	25	23.4	26.5	100	32.2	27.9	29.8	26.5	28.1	30.4	31.2
1hck__	19.7	23.6	25.5	21.9	28.2	32.7	32.2	100	37.8	27.6	26.5	29.7	30.6	27.8
1jnk__	18.5	23.2	26.7	21.8	22.7	26.5	27.9	37.8	100	27.2	23.1	26.8	28.5	29.9
1kob_a	17.8	21.6	22.4	20.8	22.4	24	29.8	27.6	27.2	100	43.1	28.1	32.5	33.5
1tki_a	14.1	18.3	20.1	19.7	19.9	25.3	26.5	26.5	23.1	43.1	100	25.4	32.7	32.4
1cdk_a	20.7	21.4	21.9	21.1	22	23.3	28.1	29.7	26.8	28.1	25.4	100	34.6	32.4
1phk__	17.6	23	19.6	20.8	22.7	26.2	30.4	30.6	28.5	32.5	32.7	34.6	100	36.3
1a06__	15.9	19.4	25.4	20	21.4	25.4	31.2	27.8	29.9	33.5	32.4	32.4	36.3	100

b)

PDB entry	1csn__	1b6c_b	1fgi_a	1ir3_a	2src__	1a6o__	1f3m_c	1hck__	1jnk__	1kob_a	1tki_a	1cdk_a	1phk__	1a06__
	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB
1csn__	0.00 0.00	3.25 3.68	3.28 3.60	3.08 3.43	3.26 3.61	2.80 3.21	3.56 3.88	2.96 3.36	3.15 3.53	3.07 3.34	3.13 3.32	2.58 2.98	2.62 2.92	3.30 3.53
1b6c_b	3.25 3.68	0.00 0.00	2.74 3.04	2.57 2.70	2.62 2.94	2.86 3.29	3.55 3.89	3.07 3.42	2.96 3.25	3.02 3.41	3.12 3.45	2.69 3.03	2.58 2.96	3.12 3.50
1fgi_a	3.28 3.60	2.74 3.04	0.00 0.00	2.04 2.25	2.77 2.96	2.95 3.36	2.98 3.34	3.02 3.24	3.07 3.38	3.07 3.42	3.03 3.37	3.35 3.66	2.96 3.19	2.57 2.83
1ir3_a	3.08 3.43	2.57 2.70	2.04 2.25	0.00 0.00	3.01 2.89	2.87 3.23	3.22 3.60	3.42 3.57	2.72 2.95	3.15 3.50	3.06 3.31	2.85 3.13	2.60 2.81	2.99 3.24
2src__	3.26 3.61	2.62 2.94	2.77 2.96	3.01 2.89	0.00 0.00	2.99 3.30	3.60 3.71	2.61 2.77	3.41 3.67	3.19 3.42	3.21 3.34	2.87 3.25	2.75 3.03	3.09 3.27
1a6o__	2.80 3.21	2.86 3.29	2.95 3.36	2.87 3.23	2.99 3.30	0.00 0.00	3.26 3.63	2.21 2.61	2.42 2.89	2.56 2.86	2.66 2.97	2.29 2.67	2.17 2.54	2.81 3.21
1f3m_c	3.56 3.88	3.55 3.89	2.98 3.34	3.22 3.60	3.60 3.71	3.26 3.63	0.00 0.00	3.37 3.55	3.43 3.66	2.54 2.83	2.57 2.79	3.46 3.69	3.07 3.16	2.59 2.79
1hck__	2.96 3.36	3.07 3.42	3.02 3.24	3.42 3.57	2.61 2.77	2.21 2.61	3.37 3.55	0.00 0.00	3.07 3.28	2.81 3.06	3.01 3.26	2.70 2.80	2.48 2.60	2.85 2.97
1jnk__	3.15 3.53	2.96 3.25	3.07 3.38	2.72 2.95	3.41 3.67	2.42 2.89	3.43 3.66	3.07 3.28	0.00 0.00	3.07 3.36	3.17 3.42	2.75 3.08	2.68 2.86	2.97 3.32
1kob_a	3.07 3.34	3.02 3.41	3.07 3.42	3.15 3.50	3.19 3.42	2.56 2.86	2.54 2.83	2.81 3.06	3.07 3.36	0.00 0.00	1.25 1.50	2.68 2.87	2.12 2.20	2.01 2.27
1tki_a	3.13 3.32	3.12 3.45	3.03 3.37	3.06 3.31	3.21 3.34	2.66 2.97	2.57 2.79	3.01 3.26	3.17 3.42	1.25 1.50	0.00 0.00	2.83 3.00	2.12 2.22	1.96 2.22
1cdk_a	2.58 2.98	2.69 3.03	3.35 3.66	2.85 3.13	2.87 3.25	2.29 2.67	3.46 3.69	2.70 2.80	2.75 3.08	2.68 2.87	2.83 3.00	0.00 0.00	1.55 1.83	2.83 2.99
1phk__	2.62 2.92	2.58 2.96	2.96 3.19	2.60 2.81	2.75 3.03	2.17 2.54	3.07 3.16	2.48 2.60	2.68 2.86	2.12 2.20	2.12 2.22	1.55 1.83	0.00 0.00	2.45 2.56
1a06__	3.30 3.53	3.12 3.50	2.57 2.83	2.99 3.24	3.09 3.27	2.81 3.21	2.59 2.79	2.85 2.97	2.97 3.32	2.01 2.27	1.96 2.22	2.83 2.99	2.45 2.56	0.00 0.00

*The entries are coloured according to the similarity between the two proteins in each pair.
Red: Sequence identity < 30%, Cα RMSD > 2.75 Å, yellow: 30% ≤ Sequence identity < 40%, 2.0 Å < Cα RMSD ≤ 2.75 Å, green: 40% ≤ Sequence identity < 50%, 1.5 Å < Cα RMSD ≤ 2.0 Å, white: Sequence identity ≥ 50%, Cα RMSD ≤ 1.5 Å.

Table 2. a) Pairwise sequence identities (%) and b) Cα and Cβ RMSD values (Å) for the eleven protein kinases in set B*.

a)

PDB entry	1byg_a	1fgk_a	1fvr_a	1iep_a	1ir3_a	1irk__	1k3a_a	1qcf_a	1qpc_a	1vr2_a	2src__
1byg_a	100	38.7	35.9	45.2	37.4	37.4	36.2	42.2	44	39.4	43.6
1fgk_a	38.7	100	40.4	38.1	36.1	36.8	36.7	35.2	36.8	53	37.1
1fvr_a	35.9	40.4	100	41.2	34.9	34.9	35.8	35.8	36.4	38.6	35.7
1iep_a	45.2	38.1	41.2	100	39.8	40.2	43	48.1	48.1	38.5	48.3
1ir3_a	37.4	36.1	34.9	39.8	100	100	80.4	36.4	37.5	38.3	38.8
1irk__	37.4	36.8	34.9	40.2	100	100	79.4	36.7	37.1	38.3	39.1
1k3a_a	36.2	36.7	35.8	43	80.4	79.4	100	35.8	37	37.3	36.7
1qcf_a	42.2	35.2	35.8	48.1	36.4	36.7	35.8	100	75.7	41.5	66.4
1qpc_a	44	36.8	36.4	48.1	37.5	37.1	37	75.7	100	40.7	66.3
1vr2_a	39.4	53	38.6	38.5	38.3	38.3	37.3	41.5	40.7	100	35.5
2src__	43.6	37.1	35.7	48.3	38.8	39.1	36.7	66.4	66.3	35.5	100

b)

PDB entry	1byg_a	1fgk_a	1fvr_a	1iep_a	1ir3_a	1irk__	1k3a_a	1qcf_a	1qpc_a	1vr2_a	2src__
	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB	CA CB
1byg_a	0.00 0.00	2.04 2.27	1.97 2.20	1.84 2.01	2.30 2.47	2.25 2.38	2.62 2.48	2.30 2.53	2.17 2.32	2.25 2.47	2.17 2.41
1fgk_a	2.04 2.27	0.00 0.00	2.12 2.50	1.84 1.93	2.21 2.39	2.03 2.17	2.17 2.23	2.87 3.24	1.88 2.12	1.25 1.45	2.85 3.19
1fvr_a	1.97 2.20	2.12 2.50	0.00 0.00	2.33 2.59	2.72 3.20	2.57 2.72	2.61 2.93	2.87 3.11	2.00 2.36	2.02 2.39	3.21 3.46
1iep_a	1.84 2.01	1.84 1.93	2.33 2.59	0.00 0.00	2.55 2.70	1.79 2.04	2.38 2.66	2.53 2.71	2.20 2.22	1.58 1.67	2.65 2.84
1ir3_a	2.30 2.47	2.21 2.39	2.72 3.20	2.55 2.70	0.00 0.00	2.63 2.66	1.14 1.20	2.52 2.66	1.83 1.91	2.56 2.57	2.58 2.71
1irk__	2.25 2.38	2.03 2.17	2.57 2.72	1.79 2.04	2.63 2.66	0.00 0.00	2.40 2.86	3.34 3.40	2.95 3.40	1.95 2.29	3.12 3.22
1k3a_a	2.62 2.48	2.17 2.23	2.61 2.93	2.38 2.66	1.14 1.20	2.40 2.86	0.00 0.00	3.07 3.27	1.68 1.76	2.18 2.17	3.12 3.30
1qcf_a	2.30 2.53	2.87 3.24	2.87 3.11	2.53 2.71	2.52 2.66	3.34 3.40	3.07 3.27	0.00 0.00	2.13 2.25	3.04 3.33	1.96 2.13
1qpc_a	2.17 2.32	1.88 2.12	2.00 2.36	2.20 2.22	1.83 1.91	2.95 3.40	1.68 1.76	2.13 2.25	0.00 0.00	2.01 2.13	2.30 2.41
1vr2_a	2.25 2.47	1.25 1.45	2.02 2.39	1.58 1.67	2.56 2.57	1.95 2.29	2.18 2.17	3.04 3.33	2.01 2.13	0.00 0.00	3.10 3.48
2src__	2.17 2.41	2.85 3.19	3.21 3.46	2.65 2.84	2.58 2.71	3.12 3.22	3.12 3.30	1.96 2.13	2.30 2.41	3.10 3.48	0.00 0.00

*The entries are coloured according to the similarity between the two proteins in each pair.
Red: Sequence identity < 30%, Cα RMSD > 2.75 Å, yellow: 30% ≤ Sequence identity < 40%, 2.0 Å < Cα RMSD ≤ 2.75 Å, green: 40% ≤ Sequence identity < 50%, 1.5 Å < Cα RMSD ≤ 2.0 Å, white: Sequence identity ≥ 50%, Cα RMSD ≤ 1.5 Å.

2.1.2 Homology model construction

A modelling pipeline has been developed for automatic all-against-all homology modelling from a multiple sequence alignment.[7] Two different homology modelling tools, WHAT IF (simple and advanced version)[32] and MODELLER,[33, 26, 1] can be used with this pipeline. This modelling pipeline has been used with the two sets of protein kinases described above. A multiple sequence alignment of the protein kinases in set A was created for another, separate research project.[34] This sequence alignment was based on a structural alignment made with the CE program,[31] and manually edited based on prior knowledge about the functionality of different regions of the protein kinase structures. A multiple sequence alignment of the protein kinases in set B was first made using ClustalX.[35] This alignment was then aligned manually to the alignment of set A. For both sets of kinase structures, homology models were constructed for each sequence in the multiple sequence alignment of the set using, in turn, each of the other structures as template. This resulted in 292 homology models, made using templates having between 14 and 80% sequence identity to the target.

The advanced version of WHAT IF (WI-advanced) was used for the homology modelling in this work. WHAT IF advanced maintains the backbone conformation of the template structure unchanged, and models side-chains using a backbone-dependent rotamer library.[36] Insertions (gaps in the sequence alignment) are not modelled, and the resulting homology models thus frequently contain structural gaps.

Phosphate groups were removed from phosphorylated tyrosine residues prior to homology modelling, and crystallographic water molecules, ligands and ions were also purged from the template structures.

2.1.3 Calculation of the homology model accuracy

The homology models were verified by comparison to the experimental structures of the targets. Two different measures of the homology model quality were used: RMSD values (separate overall Cα, Cβ and heavy atom (HA) RMSD) and difference in the inter-residue contact areas between the target X-ray structure and the model structure.

2.1.3.1 RMSD calculations

The target and template X-ray structures were superposed using the CE algorithm.[31] Separate overall Cα, Cβ and heavy atom RMSD values between targets and homology models were calculated using the rotation and translation matrices from the CE superpositioning of the target-template pair for the homology model.

2.1.3.2 Calculation of differences in the inter-residue contact areas

In this work, a 1.4 Å probe was used, along with a default set of van der Waals radii derived from the CHARMM27 force field,[37] to calculate the surface areas. Hydrogen atoms were ignored in the work presented here.

2.1.4 Generation of alignment score profiles

As a measure of the similarity between the target and template primary structures in each position in the sequence alignment, the value of the PAM250 similarity matrix[29] for that particular pair of amino acids was used. As mentioned above, separate multiple sequence alignments of each of the two sets of protein kinase structures were used for the homology modelling. In order to analyse the homology model quality for both sets of kinases simultaneously, the two sequence alignments used in the homology modelling were aligned to each other as described above. The alignment scores were generated based on this common multiple sequence alignment (shown in Figure 4). To separate non-modelled residues (caused by gaps in the sequence alignment) from modelled residues, the score value for a non-modelled residue was set to –100.

2.2 Multivariate regression analysis of the homology model quality data

PLS regression was used to analyse the homology model quality data. Cα, Cβ and heavy atom RMSD were analysed together with PLS2, while a separate PLS1 model was made for the contact area error. The data set was centred prior to the regression analysis, and random leave-ten-out cross-validation was used. No variable selection was carried out. Outliers were detected by inspection of influence plots, and removed from the analysis. The number of principal components (PCs) used was chosen by inspection of the explained Y-variation from the cross-validation.

Only cases where the target and template X-ray structures were in the same conformation (either active or inactive conformation) were considered, since a homology model made using a template structure in an active conformation can not be compared to a target structure in an inactive conformation and vice versa.

2.3 Validation of the method

The predictive ability of the regression model was validated by cross-validation, as described above. The ability of the regression coefficients to identify regions of the protein structures that are difficult to model was verified by comparison of the regression coefficient pattern with the multiple sequence alignment of the 23 kinases used to train the regression model.

In order to test whether this method can be used to detect alignment errors, an alternative alignment between two randomly chosen sequences from the multiple sequence alignment of protein structure set B, 1byg and 1fvr, was generated with ClustalX.[35] A new regression analysis of the contact area error similar to that described above was carried out, with 1byg and 1fvr kept out of the analysis. Using this alternative regression model, the contact area error for a homology model of 1fvr made using 1byg as template was predicted based on alignment scores generated from the alternative sequence alignment. The X-residuals from the prediction were calculated, and compared to the mean residuals for all homology models included in the regression analysis (± two standard deviations).

To get an idea of whether a combination of different homology models might improve the model quality, the average backbone Cα atom positions between all homology models of each protein was calculated. This resulted in an average backbone conformation for each protein, based on all homology models of that protein. For each homology model, the distance from this average model was calculated for all residue positions. To test whether the average model would perform better than the individual homology models, the average (over all residues in the model) distance of each model from the average model was correlated to the model quality.

3 Results

3.1 Data sets

3.1.1 Calculated homology model accuracy

Histograms over the obtained RMSD values and contact area differences between the homology models and the target X-ray structures are given in Figure 3.

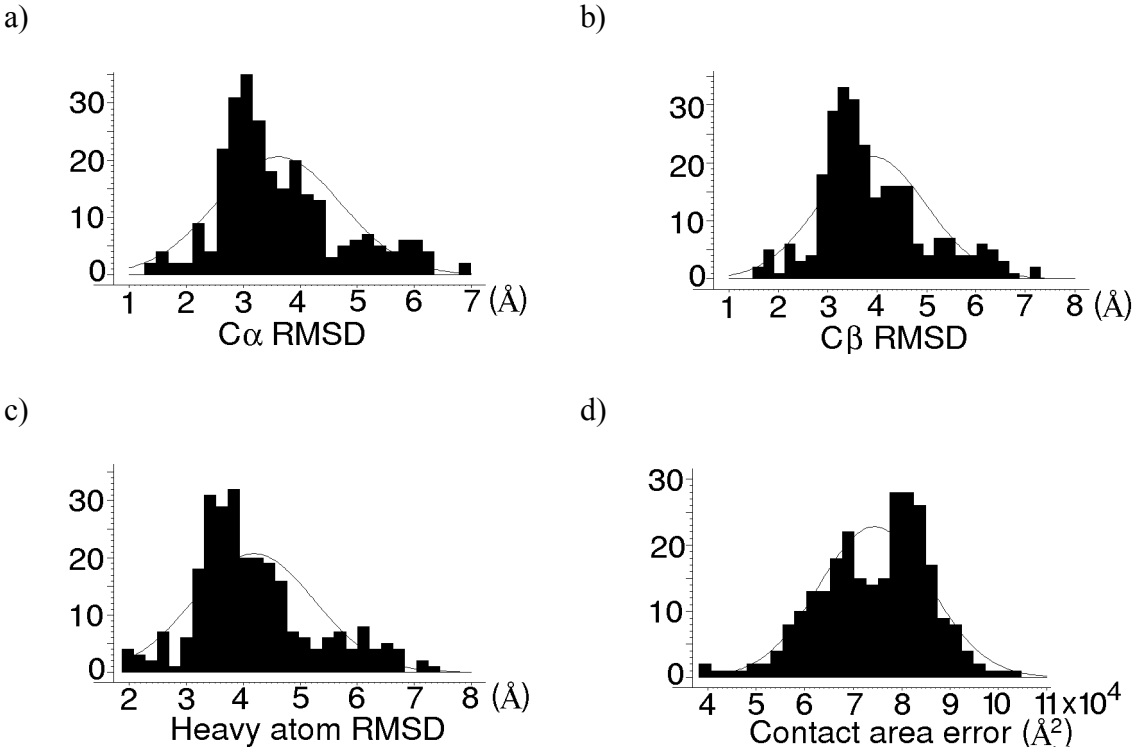


Figure 3. Histograms over a) Cα, b) Cβ and c) heavy atom RMSD values (Å) and d) contact area differences ($Å^2$) between the homology models obtained with WI-advanced and the target X-ray structures. Results for both kinase structure sets are shown in all histograms.

The histogram in Figure 3 a) shows that most of the homology models have Cα RMSD values between 2.5 and 4 Å. The relatively high RMSD values are probably caused by the wide range of sequence identities between the targets and templates used for the homology modelling.

3.1.2 Alignment score profiles

Figure 4 shows the sequence alignment that was used to generate the alignment scores for the 23 proteins studied, together with the alignment score profiles for all homology models.

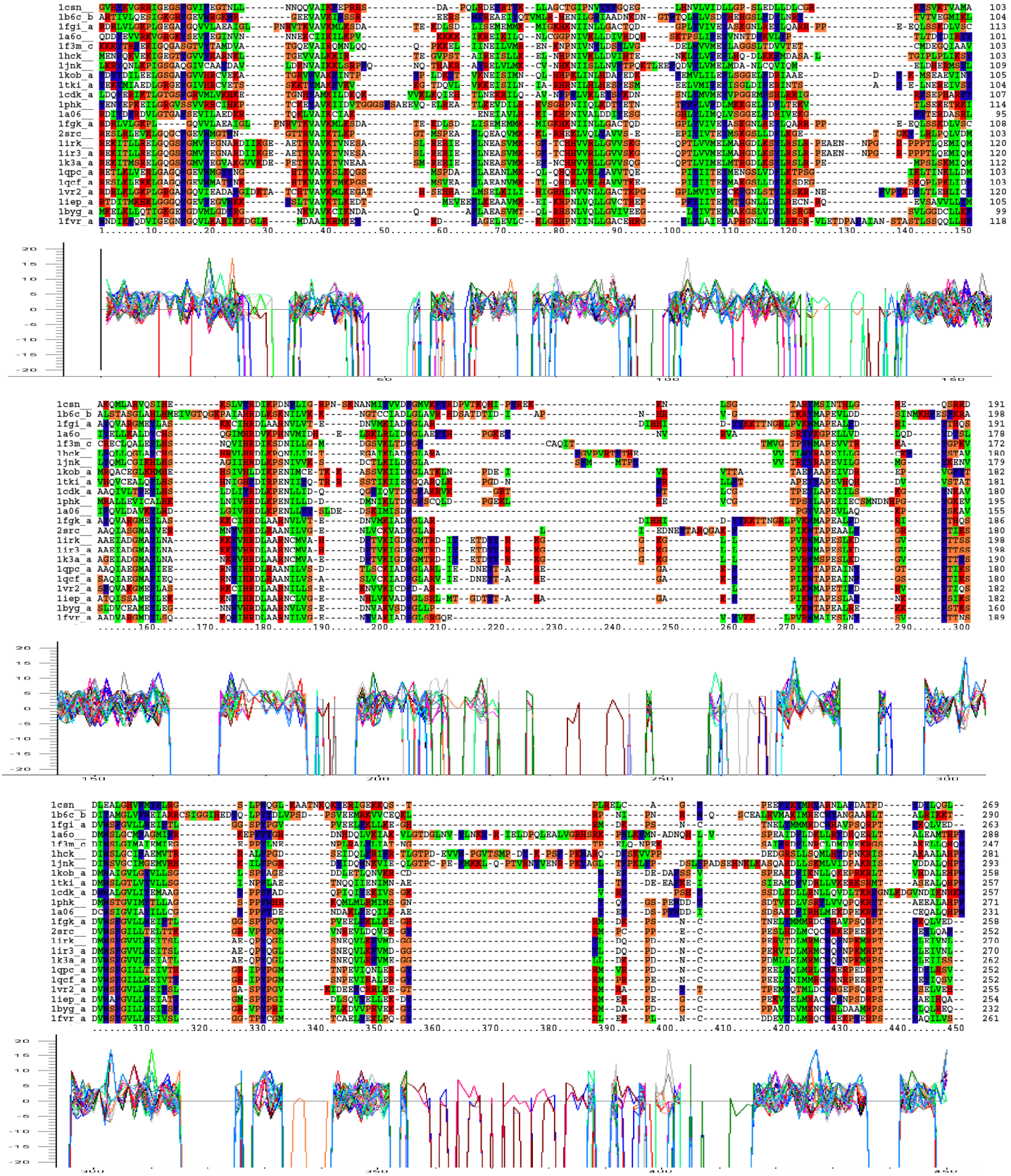


Figure 4. Multiple sequence alignment of the 23 protein kinases studied. This sequence alignment was used to generate the alignment score profiles shown below the alignment. The values on the horizontal axis correspond to the alignment positions. Score profiles for all homology models are shown.

3.2 Regression model for prediction of homology model quality

The homology model quality dataset was analysed with PLS regression. The predicted (from cross-validation) Cα, Cβ and heavy atom RMSD for the homology models are shown in Figure 5. The predicted contact area error from the cross-validation is shown in Figure 6. The PLS score plots from the regression analysis show a separation of the samples according to sequence identity between the target and template. Statistical data from the regression analysis are given in Table 3 and 4.

Table 3. Statistical data for the PLS2 regression model for Cα, Cβ and heavy atom (HA) RMSD*.

Principal components	q (Cα RMSD)	q (Cβ RMSD)	q (HA RMSD)	Explained Y-variation (%)
13	0.73	0.75	0.76	55.6

*The statistical data are from the cross-validation results.

Table 4. Statistical data for the PLS1 regression model for the contact area error*.

Principal components	q	Explained Y-variation (%)
13	0.88	77.3

*The statistical data are from the cross-validation results.

The results from this multivariate regression analysis show that the prediction of the contact area error is better than the RMSD value prediction. Inter-residue contact area errors are not affected to that extent by one single loop conformation and are not dependent on structural superposition.[7]

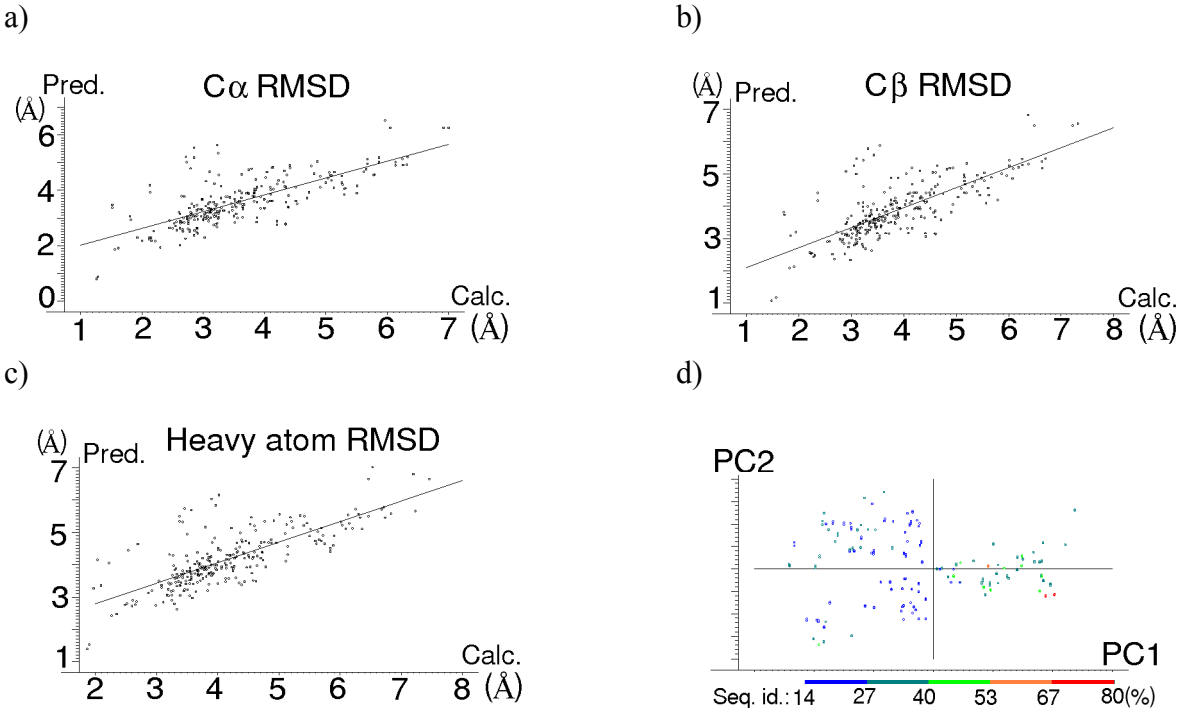


Figure 5. Predicted (from cross-validation) versus calculated values for a) Cα RMSD (Å), b) Cβ RMSD (Å) and c) heavy atom RMSD (Å) for the homology models made using WHAT IF advanced. The PLS score plot d) of PC1 vs. PC2 is also shown. The samples are coloured according to sequence identity between target and template.

a) b)

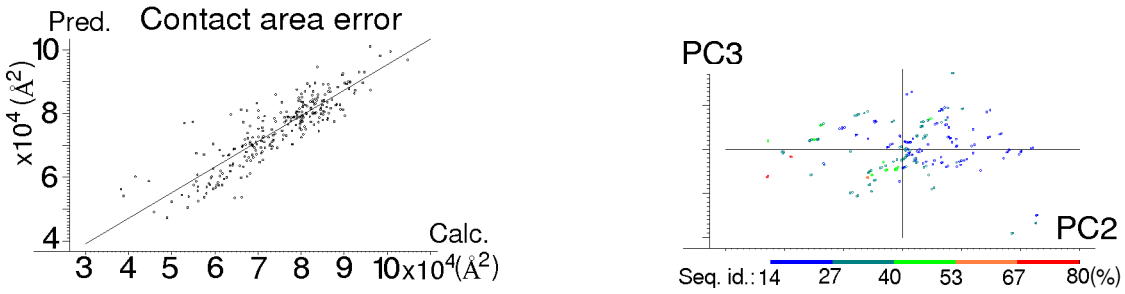


Figure 6. a) Predicted (from cross-validation) versus calculated contact area error (Å^2) for the homology models made using WHAT IF advanced. b) The PLS score plot of PC2 vs. PC3. The samples are coloured according to sequence identity between target and template.

The results presented in Figure 5 and 6 show that the homology model quality can be predicted with relatively high accuracy for the protein kinase family. Hence, the quality of future homology models can be predicted from alignment score profiles generated from substitution matrices. Similar regression models can be made for other protein families.

3.3 Validation of the method

Figure 7 shows the regression coefficients from the regression analysis. The regression coefficients from the regression model for the contact area error are shown together with the multiple sequence alignment used to generate the alignment score profiles in Figure 8. A comparison of the regression coefficients with the multiple sequence alignment shows that regions of the sequence alignment that contain many gaps (regions where the sequence alignment is of low quality) correspond to regions with large variations in the regression coefficients. Hence, the regression coefficients can be used to identify regions that are difficult to model, as well as alignment errors.

a) b)

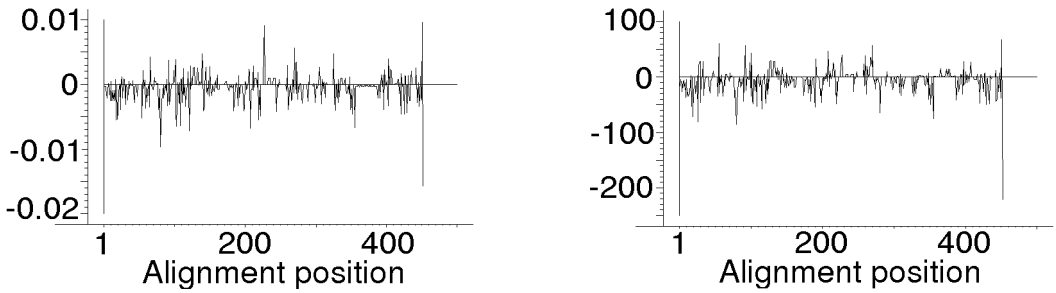


Figure 7. Regression coefficients from the regression analysis of a) the heavy atom RMSD values and b) the contact area error. The numbers on the horizontal axis correspond to the alignment positions.

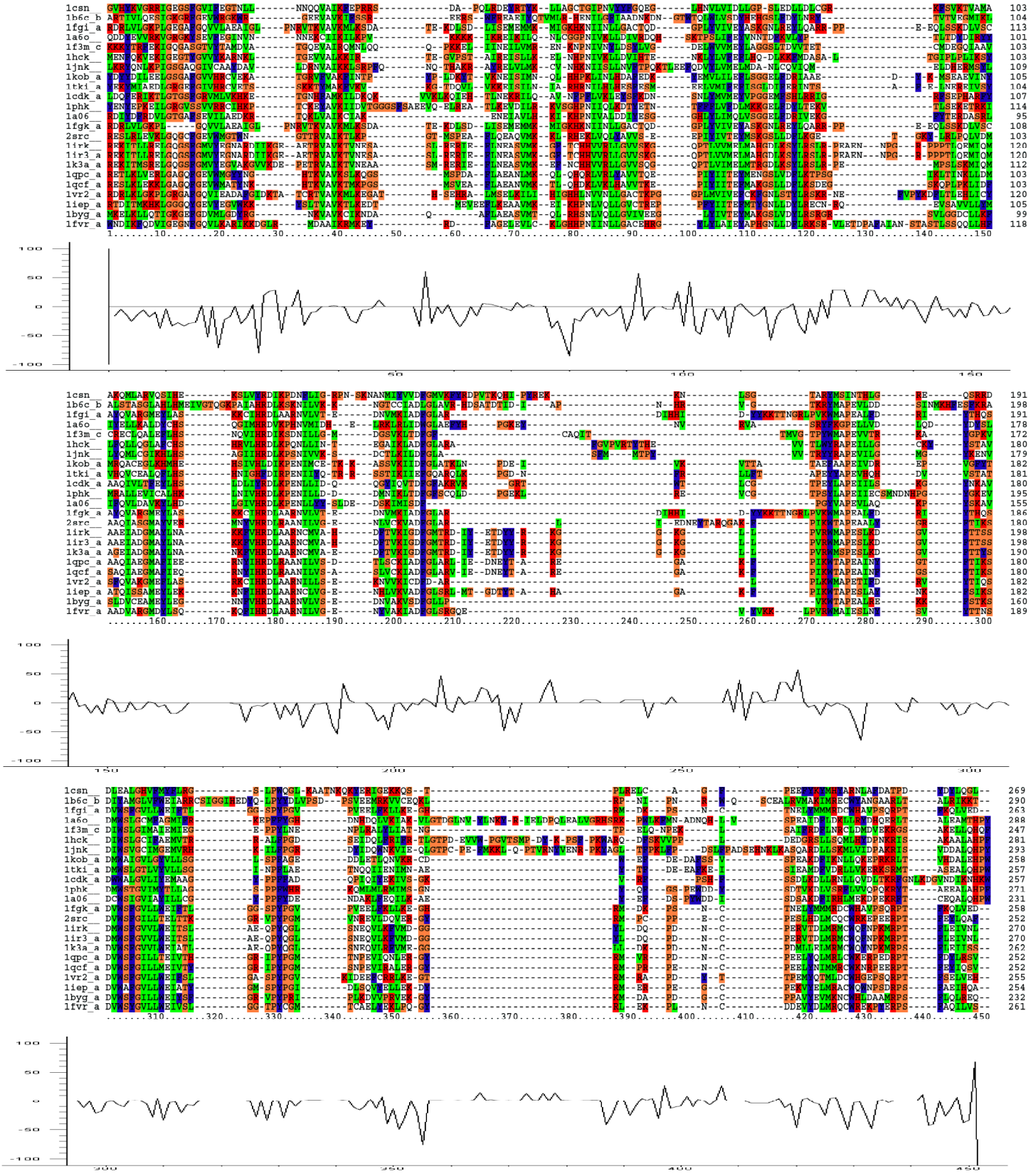


Figure 8. Regression coefficients from the regression analysis of the contact area error shown together with the multiple sequence alignment used to generate the alignment score profiles. The numbers on the horizontal axis correspond to the alignment positions.

The residuals from prediction of the model quality for new homology models can be used to identify proteins that are difficult to model with homology modelling due to large deviations from the other members of the protein family. As explained earlier, such outliers can be identified by

inspection of influence plots. The influence plots in Figure 9 show that no outliers that have a large effect on the results are present. The kinase structures with PDB entries 1f3m and 1b6c were previously removed from the PLS regression analysis because they were outliers.

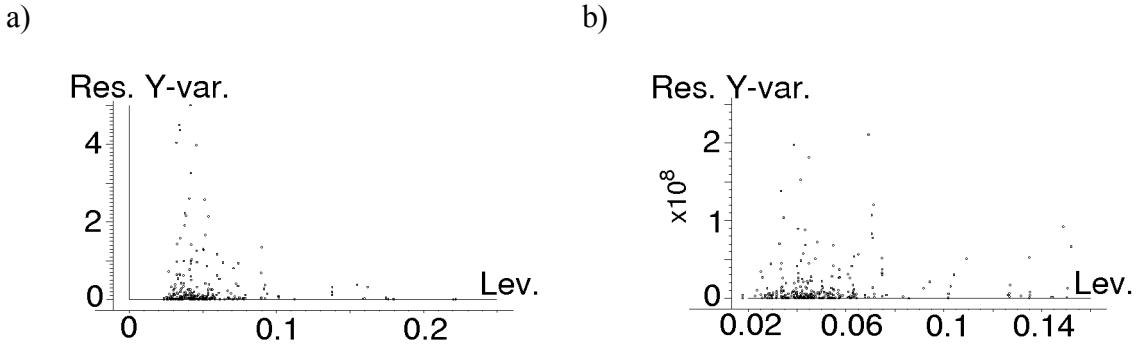


Figure 9. Influence plots (residual Y-variance versus leverage) from the regression analysis of a) the RMSD values and b) the contact area error.

Comparison of the residuals (for each alignment position) from prediction for a new homology model to the residuals for the homology models included in the regression analysis can e.g. reveal errors in the sequence alignment. Such alignment errors will lead to deviations in the residual pattern. To test this hypothesis, an alternative alignment between 1byg and 1fvr was generated. This alignment is shown together with the original alignment in Figure 10. Comparison of these two alignments reveals that the new alignment contains several deviations from the original alignment. Since the original alignment was corrected based on prior knowledge about the functionality of protein kinases, that alignment is more likely to be correct than the alternative one. Based on the alternative alignment, new alignment scores were generated, and the X-residuals from prediction of the contact area error were calculated using a regression model that had not been trained on 1byg and 1fvr. These residuals were compared to the residuals for all homology models for which the regression model was trained. The results are shown in Figure 10. Comparison of the two alignments of 1byg and 1fvr and the curves in Figure 10 shows that in regions where the two alignments differ, the residuals for 1byg and 1fvr have large deviations from the mean residuals for the homology models included in the regression analysis. Hence, the X-residuals can provide useful information about alignment errors. As seen from Figure 10, there are a couple of regions where the residuals for 1byg and 1fvr have large deviations from the mean residuals even though the alignments are identical. Hence, such deviations from the mean can also be caused by other factors than alignment errors, and can only be used to identify regions where a closer look at the alignment might be necessary.

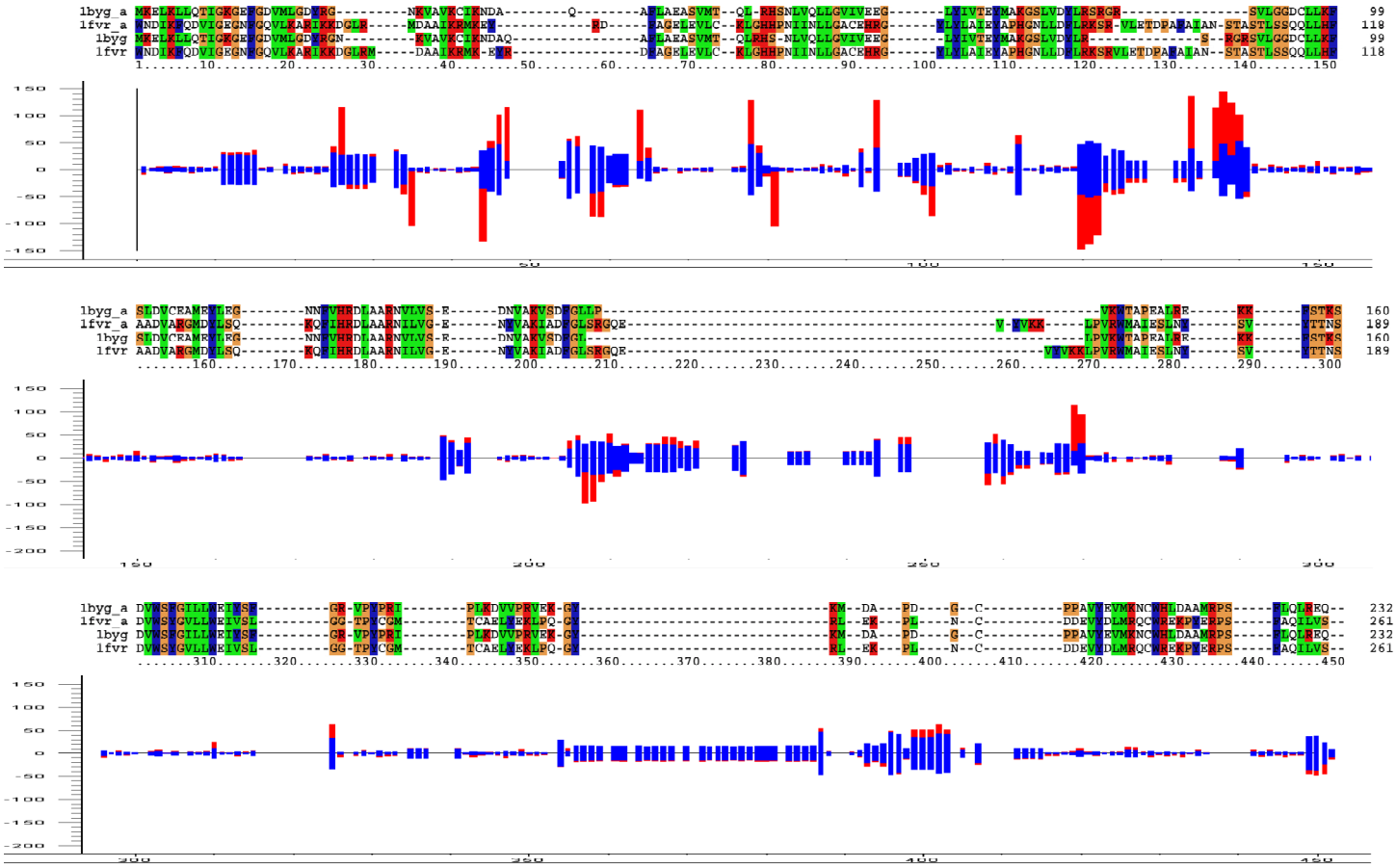


Figure 10. Alternative alignment between 1byg and 1fvr (last two sequences), aligned to the original alignment (first two sequences) from the multiple sequence alignment in Figure 4. The X-residuals from the prediction of the contact area error based on the alternative alignment of 1byg and 1fvr (red curve) are compared to the mean X-residuals for all homology models included in the regression analysis ± two standard deviations (blue curve). The numbers on the horizontal axis correspond to the alignment positions.

To test whether an average model would perform better than the individual homology models, the average (over all residues in the homology model) distance of each homology model from the average model was correlated to the model quality. Only the data for the proteins in kinase structure set B (sequence identities of 35-80%) gave meaningful results. The results are shown in Figure 11.

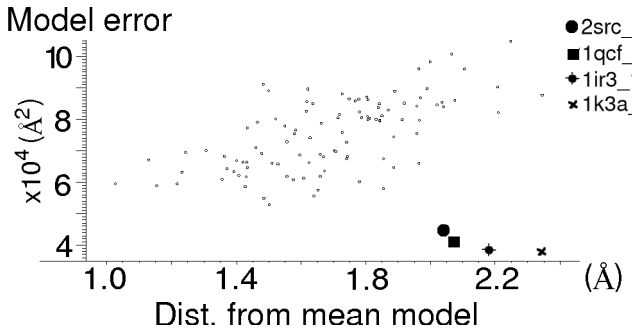


Figure 11. Contact area error (Å^2) for the homology models of the proteins in kinase structure set B versus the average (over all residues in the homology model) distance of each homology model from the average backbone conformation (Å) (2src_1qcf means the homology model of 2src made using 1qcf as template, and likewise for the other homology models).

The fact that the homology model quality is correlated with the distance from the average backbone conformation indicates that using a combination of several homology models might improve the model quality. Keeping the four marked outliers in Figure 11 out gives a correlation of 0.64

between the model error and the average distance from the mean model. Generation of plots like the one shown in Figure 11 can be used to identify cases where a single template performs better than a combination of several templates. An example of a target-template pair where using a single template gives the best result is 1k3a and 1ir3. These two structures are so similar, that using multiple templates in combination would probably introduce errors to the homology model. The sequence identity between 1k3a and 1ir3 is 80.4%. Such target-template pairs are placed in the lower, right hand part of the plot in Figure 11, since the homology model quality will be high even though the homology model differs a lot from the average model. Hence, when the similarity between the target and template structures is high, using a single template is probably better than using an average model, since the template structure is more similar to the target than an average model will be. The other templates will make the average model differ more from the target.

4 Discussion

The method presented here provides a new way to predict the quality of homology models directly from the sequence alignment between the target and template sequences. This method can be used prior to the actual homology model generation. This is new, since existing methods for model quality prediction work on the protein structure models. Hence, the time spent generating the homology models can be saved by using this method to rule out cases in which homology modelling is likely to fail, and when it may succeed. The correct templates to use for the homology modelling can thereby more effectively be found. Since separate regression models can be made for different protein families and different homology modelling methods, homology model quality prediction can also guide the choice of modelling method.

Combination of several template structures in the homology modelling is widely used. The underlying idea is that multiple template structures provide more information than a single structure does. If the correct template structures are chosen, this is probably true. However, including structural information from template structures that do not have the required similarity to the target may introduce errors in the final homology model. In this case, using a single template with high similarity to the target is better than using this template in combination with other templates of lower similarity. The method presented here can be used to find the optimal combination of templates, and in which cases using a single template may give the best result.

In some cases it is best to model different domains of the protein structure separately. Homology model quality prediction is useful for identifying what domains to model separately, and what templates to use for the different domains. Plots of the regression coefficients from the regression analysis can be used to identify regions that are difficult to model, and X-residuals from the prediction can be used to detect alignment errors. Influence plots can be used to detect members of the protein family that will be difficult to model due to large deviations from the other members of the family.

One problem with most homology modelling methods that use a combination of multiple template structures is that a primary template (typically the one having the highest sequence identity to the target) is chosen, and information from the other template structures is often only used in gap regions. This makes the homology model very dependent on the primary template, and often this results in a model that is more similar to the primary template than to the target. This is, however, only the case when there are errors in the target-template alignment used for the homology modelling.[1] Hence, this technique is most useful in cases where a template structure of relatively high sequence identity is available. It is also difficult to obtain a correct sequence alignment in cases where the templates have low sequence identity to each other and to the target. In cases where only template structures of low sequence identity are available, including structural information from all templates along the entire sequence might be better than choosing one of them as a primary template. A reasonable question is therefore: Is it possible to combine several homology models of

low overall quality by using e.g. a weighted average of the backbone positions for each residue? The weights for each homology model should vary according to the similarity to the target sequence in that region. In this way, each homology model would contribute differently in different regions according to the local similarity to the target. One way to weight this average would be to use alignment score profiles like those generated here. Different homology modelling methods might also perform differently in different regions of the protein. Hence, a combination of homology models generated using several modelling methods might improve the model quality. One problem with this procedure is that averaging the side-chain positions does not make sense. Hence, the side-chain conformations have to be determined after the backbone average is calculated.

5 Conclusions

A new method for prediction of homology model quality has been presented, which is a useful tool both for selection of template structures for the homology modelling, and for detection of alignment errors. This method can also be used to identify problem regions of a protein structure, as well as proteins that are difficult to model with homology modelling due to large deviations from the other members of the protein family. It will also be a useful tool for improving the homology model quality by combination of several homology models. This method has been applied to protein kinases, and can easily be extended to other protein families.

Acknowledgements

Thanks to Eric Scheeff and Philip E. Bourne at San Diego Supercomputer Center for providing the multiple sequence alignment of the protein kinases in set A. Michael Gribskov at San Diego Supercomputer Center is thanked for helping with the alignment of the kinases in set B. Thanks to Jens E. Nielsen for providing the homology modelling pipeline, and to Stewart Adcock for helping with the calculations of the inter-residue contact areas. Thanks also to Prof. J. Andrew McCammon and his research group at the Department of Chemistry and Biochemistry at University of California, San Diego, for allowing me to visit their group and carry out the homology modelling work there.

Dr. Finn Drabløs at the Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology and Endre Anderssen at the Department of Chemistry, Norwegian University of Science and Technology are thanked for helpful discussions.

The Norwegian Research Council is thanked for financial support.

References

1. Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291-325.
2. Baker, D.; Sali, A. Protein Structure Prediction and Structural Genomics. *Science* **2001**, *294*, 93-96.
3. Al Lazikani, B.; Jung, J.; Xiang, Z.; Honig, B. Protein structure prediction. *Curr. Opin. Chem. Biol.* **2001**, *5*, 51-56.
4. Schonbrun, J.; Wedemeyer, W. J.; Baker, D. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **2002**, *12*, 348-354.
5. Qian, B.; Goldstein, R. A. Optimization of a new score function for the generation of accurate alignments. *Prot. Struct. Func. Gen.* **2002**, *48*, 605-610.
6. Koehl, P.; Levitt, M. A brighter future for protein structure prediction. *Nat. Struct. Biol.* **1999**, *6*, 108-111.
7. Liu, J.; Tøndel, K.; Adcock, S.; Gribskov, M.; Niedner, H. R.; McCammon, J. A.; Nielsen, J. E. Homology Modelling of Protein Kinases. *Unpublished results.*
8. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystalogr.* **1993**, *26*, 283-291.
9. Laskowski, R.A.; Rullmann, J. A. C.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **1996**, *8*, 477-486.
10. Oldfield, T. J. Squid: a program for the analysis and display of data from crystallography and molecular dynamics. *J. Mol. Graphics* **1992**, *10*, 247-252.
11. Hooft, R. W. W.; Sander, C.; Vriend, G. Verification of protein structures: side-chain planarity. *J. Appl. Crystallogr.* **1996**, *29*, 714-716.
12. Lüthy, R.; Bowie, J. U.; Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **1992**, *356*, 83-85.
13. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Prot. Struct. Func. Gen.* **1993**, *17*, 355-362.
14. Topham, C. M.; Srinivasan, N.; Thorpe, C. J.; Overington, J. P.; Kalsheker, N. A. Comparative modelling of major house dust mite allergen der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng.* **1994**, *7*, 869-894.
15. Melo, F.; Feytmans, E. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **1998**, *277*, 1141-1152.
16. Venclovas, C.; Zemla, A.; Fidelis, K.; Moult, J. Criteria for evaluating protein structures derived from comparative modeling. *Prot. Struct. Func. Gen.* **1997**, *Suppl. 1*, 7-13.
17. Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP): Round IV. *Prot. Struct. Func. Gen.* **2001**, *Suppl. 5*, 2-7.
18. Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Prot. Struct. Func. Gen.* **2003**, *53, Suppl. 6*, 334-339.
19. Cristobal, S.; Zemla, A.; Fischer, D.; Rychlewski, L.; Elofsson, A. A study of quality measures for protein threading models. *BMC Bioinformatics* **2001**, *2, 5*.
20. Abagyan, R. A.; Totrov, M. M. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **1997**, *268*, 678-685.
21. le Grand, S. M.; Merz Jr., K. M. Rapid Approximation to Molecular Surface Area via the use of Boolean logic and look-up tables. *J. Comp. Chem.* **1993**, *14*, 349-352.
22. Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Prot. Struct. Func. Gen.* **1994**, *18*, 309-317.
23. Russell, R. B.; Barton, G. J. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* **1994**, *244*, 332-350.

24. Gou, Z. Y.; Thirumalai, D. Kinetics of Protein Folding: Nucleation mechanism, time scales and pathways. *Biopolymers* **1995**, *36*, 83-102.
25. Lesk, A. M.; Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **1980**, *136*, 225-270.
26. Fiser, A.; Do, R. K.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9*, 1753-1773.
27. Mehler, E. L.; Periole, W.; Hassan, S. A.; Weinstein, H. Key issues in the computational simulation of GPCR function: representation of loop domains. *J. Comput. Aided Mol. Des.* **2002**, *16*, 841-853.
28. Wieman, H.; Tøndel, K.; Anderssen, E.; Drabløs, F. Homology-based modelling of targets for rational drug design. *Mini-Reviews in Medicinal Chemistry* **2004**, *In Press.*
29. Schwartz, R. M.; Dayhoff, M. O. Origins of Prokaryotes, Eukaryotes, Mitochondria, and Chloroplasts. A perspective is derived from protein and nucleic acid sequence data. *Science* **1978**, *199*, 395-403.
30. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.
31. Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739-747.
32. Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8*, 52-56.
33. Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779-815.
34. Scheeff, E. D.; Bourne, P. E. Evolution of the Protein Kinase-like Superfamily, from a Structural Perspective. *Unpublished results.*
35. Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **1997**, *25*, 4876-4882.
36. Chinea, G.; Padron, G.; Hooft, R. W.; Sander, C.; Vriend, G. The use of position-specific rotamers in model building by homology. Proteins **1995**, *23*, 415-421.
37. MacKerell, A. D.; Brooks, B.; Brook III, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. In *Encyclopedia of Computational Chemistry*; Schleyer, P. R., Ed.; John Wiley & Sons: New York, 1998; pp. 271-277.

Paper II

Journal of Computer-Aided Molecular Design, **16:** 831–840, 2002.
KLUWER/ESCOM

Protein Alpha Shape Similarity Analysis (PASSA): A new method for mapping protein binding sites. Application in the design of a selective inhibitor of Tyrosine kinase 2

Kristin Tøndel[a,*], Endre Anderssen[a] & Finn Drabløs[b]
[a]*Department of Chemistry, Norwegian University of Science and Technology, Sem Selands v. 14, N-7491 Trondheim, Norway;* [b]*SINTEF Unimed MR-center, N-7465 Trondheim, Norway*

Received 1 October 2002; Accepted 6 January 2003

Key words: 3D QSAR, alpha shapes, drug design, homology modelling, PLS, proteins, selectivity, tyrosine kinase inhibitors.

Summary

We have developed a method that we have called Protein Alpha Shape Similarity Analysis (PASSA), that identifies interaction sites that can be utilised to achieve selectivity towards a protein. We have shown that this method is able to identify residues of tyrosine kinases that interact with known selective inhibitors using the following test cases: Abelson (Abl) kinase in complex with STI-571 and Janus kinase 2 (Jak2) in complex with AG-490. The 3D structures of the tyrosine kinase domains of Tyrosine kinase 2 (Tyk2) and Jak2 have been predicted by homology modelling. Computational docking of AG-490 and a set of tyrphostins known not to inhibit Jak2 indicated that our homology models are able to separate inhibitors from non-inhibitors. PASSA has also been used to identify unique properties of Tyk2. According to our results, interactions with hydrogen acceptors and donors on the following residues can be utilised to achieve selectivity towards Tyk2: Y955, E1053, D1062 and S1063. These residues are placed close to non-conserved hydrophobic pockets. The PASSA results, together with results from Multiple Copy Simultaneous Search (MCSS) were used to suggest functional groups of a selective Tyk2 inhibitor.

Introduction

Protein kinases contribute to regulation and coordination of e.g. metabolism, gene expression, cell growth, cell motility, cell differentiation and cell division [1]. The Janus kinase (Jak) family of non-receptor tyrosine kinases consists of four known mammalian proteins (Jak1-3 and Tyk2) that play a critical role in initiating signalling cascades of a large number of cytokine receptors [2–5]. Tyrosine kinases are usually regulated by phosphorylation of tyrosine residues in the activation loop, located between the conserved DFG and APE motifs [6]. This tyrosine phosphorylation causes conformational changes in the activation loop, that allow ATP and protein substrates to access the active site [7].

*To whom correspondence should be addressed.
E-mail: kristito@phys.chem.ntnu.no

The Jaks catalyse phosphorylation of the Signal Transducers and Activators of Transcription (STAT) family of transcription factors [6]. After phosphorylation on tyrosine residues, the STAT molecules form homo- or heterodimers [8], which are translocated into the nucleus. The STAT proteins then bind to DNA, and activate gene transcription [2]. The Jak-STAT signalling cascade has been shown to contribute to growth and survival of e.g. human multiple myeloma cells [9], acute lymphoblastic leukaemia [10] and a variety of other malignancies [11,12]. This makes the Jaks potential targets for new cancer therapies. One way to block the function of the Jaks is to inhibit ATP binding. ATP competitive inhibitors are generally non-selective, but the development of inhibitors like STI-571 [13] shows that ATP binding sites can be used as targets for selective drugs. Since none of the Jaks have experimentally determined 3D structures at the

present time [14,15], we have made homology models [16] of the tyrosine kinase domains of Tyk2 and Jak2. A selective inhibitor of Jak2 has been reported [10]. We have therefore focused our design work on Tyk2, and used the model of Jak2 for method testing.

The quality of homology models is highly dependent on the choice of template structures. According to Chothia and Lesk [17], templates with sequence identity $> 50\%$ to the target proteins are likely to give reasonable models. If the sequence identity drops to 20%, there will be large structural differences. However, the active sites of distantly related proteins can have very similar geometries [18,19]. A weakness of using structures predicted by homology modelling as basis for the design of selective drugs is that to achieve selectivity one has to utilise variable regions of the proteins. These are the regions predicted with the lowest reliability by homology modelling techniques [20]. The ability of our homology models to separate inhibitors from non-inhibitors was tested by computational docking of the Jak2 inhibitor AG-490 and 10 other tyrphostins known not to inhibit Jak2 [10].

To design a selective inhibitor for Tyk2, we need to identify interaction sites that can form the basis for selectivity. We have developed a method that we have called Protein Alpha Shape Similarity Analysis (PASSA), which identifies residues that are unique to one protein compared to several others. A number of methods exist to map protein binding sites. Some force field based methods, such as GRID [21] and Multiple Copy Simultaneous Search (MCSS) [22], use calculated interaction energies between probe molecules and the protein. In GRID the interactions are estimated by placing a probe atom at a number of fixed grid points in the protein. MCSS does not use a fixed grid. Instead, a geometry optimisation is performed on a large number of probe molecules placed randomly in the binding site. The probe molecules that bind strongly to the protein can then be taken as a basis for placement of functional groups in e.g. combinatorial library design. Other methods use the shape of the protein to find potential binding sites, without energy calculations. An example is alpha sphere-based methods [23]. Alpha spheres are geometrical representations of protein cavities. Alpha sphere centres are often found close to atoms of docked ligands [24].

In order to determine which sites can contribute to selectivity, a number of proteins must be mapped and the binding sites compared. GRID has been used to reveal structural differences between proteins [21]. The proteins are aligned prior to GRID calculations, and the data is analysed by Principal Component Analysis (PCA) [25]. MCSS and methods using alpha shapes are less suitable for direct comparison of proteins, because of the free movement of probe molecules and the absence of a fixed frame of reference, such as a grid. Force field methods can give spurious results due to errors in alignment or structures. The Lennard-Jones- and electrostatic terms of force fields are very steep close to atomic nuclei. Small changes in atomic positions can therefore lead to large changes in the calculated energy.

To avoid some of the problems mentioned above, we have combined Gaussian property distributions (similar to those used in Comparative Molecular Similarity Index Analysis (CoMSIA) [26]) and alpha shapes [23] in order to compare proteins. The value of a property field at each point of a grid on the aligned proteins is computed as a weighted sum of property Gaussians. The resulting similarity fields can be analysed by e.g. PCA or Discriminant Partial Least Squares (DPLS) regression [25]. In DPLS, the response matrix contains binary indicator variables indicating memberships in different classes. When the purpose is to differentiate between known classes of proteins, DPLS has the advantage that the most relevant differences are extracted and summarised as a single vector of regression coefficients. The regression coefficients can be displayed as isosurfaces on the 3D structures of the proteins. Such visualisations can be combined with the results from MCSS. This combines information about binding of functional groups and potential for selectivity.

STI-571 is a selective inhibitor of Abelson (Abl) kinase, platelet-derived growth factor (PDGF) receptor and c-kit [13,27]. We have tested whether the residues identified by PASSA to be unique to Abl kinase match the residues that interact with STI-571. The same test was carried out using the homology model of Jak2 in complex with the lowest docking energy conformation of AG-490. AG-490 has been reported to also inhibit Jak1 [28] and Jak3 [11,29], but not Tyk2 [30]. This test is valid, since neither Jak1 nor Jak3 were included in this analysis. We have utilised PASSA to identify unique properties of the Tyk2 tyrosine kinase domain. The results from this analysis, together with results from MCSS runs, have been used to suggest positioning of functional groups of a selective Tyk2 inhibitor.

Methods

Homology modelling

Homology models for the kinase domains of Tyk2 (F892-Q1177) and Jak2 (S833-N1129) were made using five templates simultaneously in SwissModel [31–34]. Suitable templates were found using the SwissModel Blast search [31,32]. In cases where more than one file were available in the RCSB Protein Data Bank (PDB) [14,15] for the same protein, the structure with the best resolution was used. All templates have sequence identities of 35%–45% to the target. SwissModel estimates the model reliability (Model B-factor) [32] for each atom in the model, based on the similarity between the target protein and the templates.

Hydrogens were added to the model structures in Molecular Operating Environment (MOE^{TM}) [35], and the structures were energy minimised to an RMS gradient of 0.01 using the AMBER94 force field [36] with a smooth non-bonded cut-off of 10–12 Å. The calculations were carried out in vacuum, using a distance-dependent dielectric to approximate the solvent screening effects. The energy minimisations were performed with fixed positions for backbone atoms of high reliability regions and heavy atoms of residues in inter-domain contact regions, because it is generally known that with extensive refinement, homology models tend to get worse [37]. In domain modelling, positions of atoms forming an interface to a missing domain should be fixed during energy minimisation. Free movement in these regions can lead to sidechain conformations that are preferable energetically, but not possible in the real protein structure because of interactions with the missing parts of the protein. The quality of the structures was verified by WHAT_CHECK [38,39]. The residues for which the WHAT_CHECK routines reported unusual conformations were relaxed, and the models were again optimised to an RMS gradient of 0.01.

For comparison, three homology models were made for both Tyk2 and Jak2 using only one template for each model in SwissModel. The hydrogens of the structures were optimised as described above. All non-hydrogens were held in fixed positions during this optimisation. A structure superpositioning of the α-carbons of these models was carried out in Swiss-PdbViewer [31,40], and the Cα Root Mean Square Distance (RMSD) was calculated.

Multiple Copy Simultaneous Search (MCSS)

We have carried out MCSS [22] runs in MOE to identify binding sites for acetamide, acetaldehyde, water, methane and benzene in the ATP binding pockets of Tyk2 and Jak2. MMFF94 [41] with implicit solvent electrostatic corrections [42–44] were used for the energy minimisations. 500 copies of each probe molecule were used.

Docking analysis

In addition to WHAT_CHECK verification, the model quality was tested by computational docking. All docking calculations were carried out using Tabu search [45] in MOE [35]. Tabu search is a stochastic searching algorithm that maintains a list of previously visited conformations, to guide the searching towards better conformations. The MMFF94 force field [41] was used, and the calculations were done in vacuum, with a distance-dependent dielectric and a smooth non-bonded cut-off of 10–12 Å. MMFF94 was chosen because it has more parameters for small molecules [46] than AMBER94. The AMBER force fields are more suited for calculations on proteins. Grid-based potential fields [35] were used for estimation of the interaction energies. Hence, the potential energy grids were calculated only at the beginning of the docking procedure.

The model of Tyk2 was aligned in MOE to the structure of human cyclin-dependent kinase 2 (CDK2) present in PDB [14,15] entry 1HCK. Sequence alignments were carried out using a modified version of the Needleman and Wunsch approach [47] with a structural correction and the Gonnet similarity matrix [48]. The 3D structures were superposed as described in [49]. ATP was docked 10 runs of 25 000 iterations using the experimentally determined ATP structure present in 1HCK as the starting position.

All fragments from the MCSS run on Jak2 having negative interaction energies with the receptor were combined into a ‘molecular cluster’. The structures of the tyrphostins were aligned onto this cluster using flexible alignment [50] in MOE. This generated a set of starting conformations for each tyrphostin. All starting conformations were docked in the homology model of Jak2, using the same docking box ($150\times150\times150$ grid points with 0.15 Å spacing). 1000 iterations were used for each starting conformation.

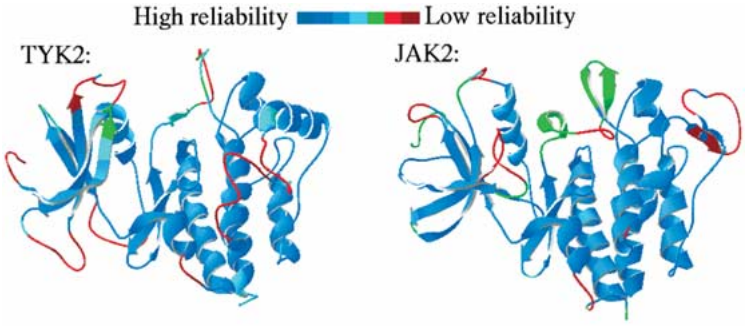


Figure 1. Predicted structures of the tyrosine kinase domains of Tyk2 and Jak2. The residues are coloured according to the Model B-factor [32] computed by SwissModel.

PASS Analysis

Property Gaussians were used to analyse the protein structures. Alpha-spheres [23] were used to identify potential hydrophobic and hydrophilic binding sites. 'Dummy' atoms placed at each alpha-sphere centre were assigned similarity field weights (ω) of +1 for either the hydrophobic or the hydrophilic field. In order to include steric effects, protein atoms were assigned a field weight of −1 for both fields (Equation (1)).

$$F(q, j) = \sum_{i=1}^{n} \frac{\omega_{ik}}{(\sigma_i \sqrt{2\pi})^3} \cdot e^{\frac{-r_{iq}^2}{2\sigma_i^2}} \qquad (1)$$

F is the value of the similarity field in grid point q of molecule j, ω_{ik} is the value of the physicochemical property k of atom i, r_{iq} is the distance between grid point q and atom i and σ_i corresponds to the atomic radii of atom i. PASSA works as follows:

1. Structural alignment of the proteins.
2. Placement of a grid surrounding the active sites of the proteins.
3. Determination of alpha-sphere positions.
4. Calculation of the value of the molecular similarity field in each grid point.
5. The molecular similarity field is taken as input to DPLS regression [25].

The regression coefficients from the DPLS regression can be mapped back onto the grid. This gives us the opportunity to visualise the regions having properties that are unique to a specific protein. All scripts were written in Scientific Vector Language (SVL) [35], and are available from the authors upon request.

The five- and one-template models of Tyk2 and Jak2, the structures used as templates in the homology modelling and PDB entries 1JST, 3LCK, 1VR2, 1IRK, 2SRC, 1AD5 and 1IEP were included in the PASS Analysis. Indicator variables for the following classes were used: Tyk2-models, Jak2-models, structures of non-Janus kinases and Abl kinase structures. The structures were superposed in MOE [35] using the same approach as mentioned in the previous section. A 3D grid was centred at the nucleotide-binding loop of Tyk2 (L903-V911). 50×50×40 grid points were used with a grid spacing of 0.75 Å. The DPLS regression was done in MATLAB™ [51], using the PLS Toolbox® [52]. All columns with standard deviation $< 10^{-4}$, or three or less nonzero entries were removed from the PASSA data. The data for each physicochemical property was set to equal variation by dividing each data point by the sum of the singular values of the data for this property. The resulting matrix was used as regressor data in the DPLS regression.

Table 1. Docking energies for the 11 tyrphostins used for verification of the model quality[a]

Tyrphostin	Docking energy (kJ/mol)
AG-490	−32.75
AG-30	−23.35
AG-18	−17.47
AG-126	11.19
AG-1295	19.73
AG-294	34.20
AG-370	40.29
AG-1112	47.23
AG-1007	56.52
AG-1478	69.41
AG-879	101.44

[a]AG-490 is an inhibitor of Jak2 [10], while all the other tyrphostins are non-active.
AG-1007 resembles AG-490 in structure (Figure 4).

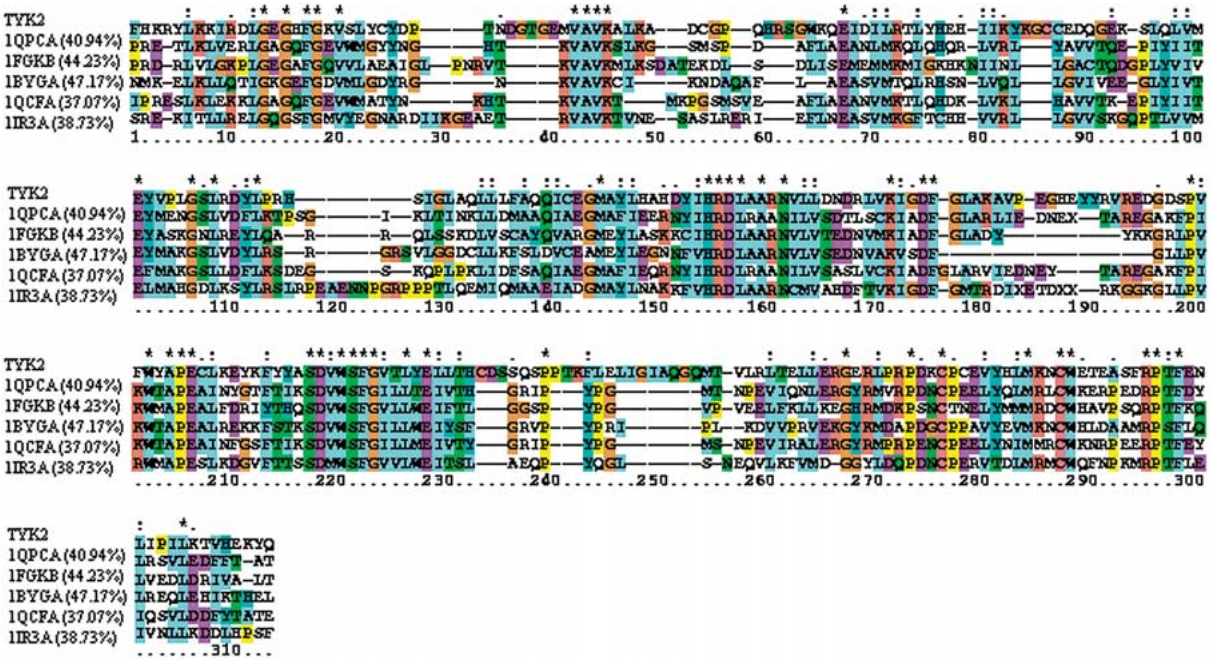


Figure 2. SwissModel multiple sequence alignment of Tyk2 (F892-Q1177) with the templates used for homology modelling. The sequence identities between the templates and the Tyk2 tyrosine kinase domain are shown together with the PDB [14,15] identifiers. The nucleotide-binding loop and the DFG motif are given in alignment positions 13–46 and 176–179, respectively.

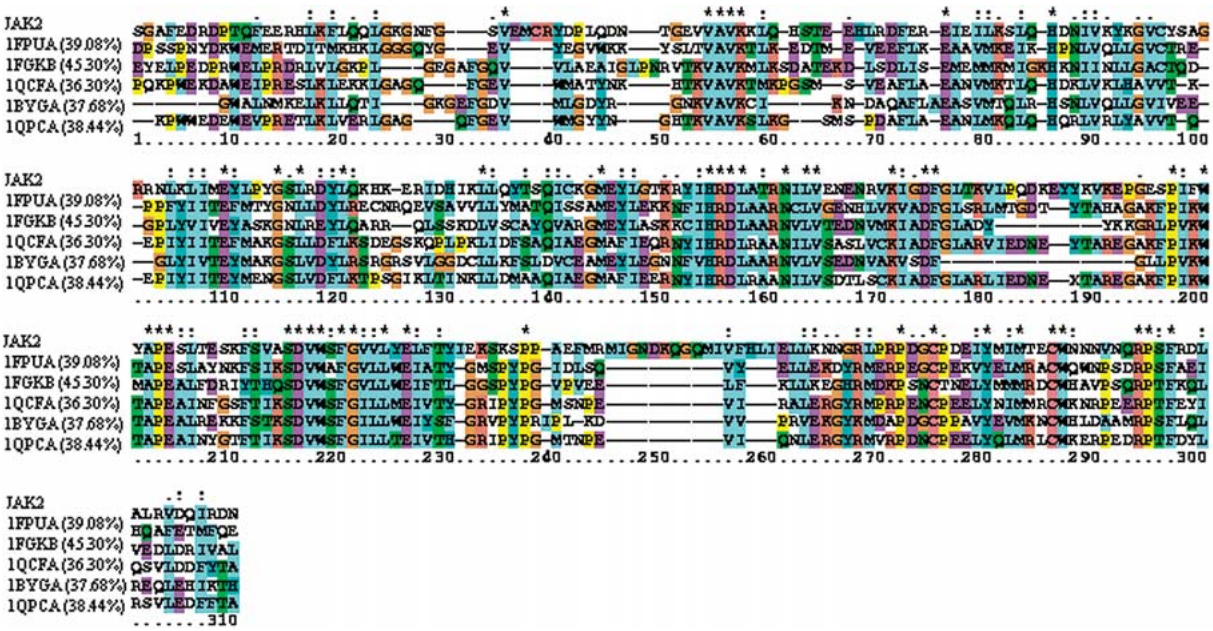


Figure 3. SwissModel multiple sequence alignment of Jak2 (S833-N1129) with the templates used for homology modelling. The sequence identities between the templates and the Jak2 tyrosine kinase domain are shown together with the PDB [14,15] identifiers. The nucleotide-binding loop and the DFG motif are given in alignment positions 24–58 and 176–178, respectively.

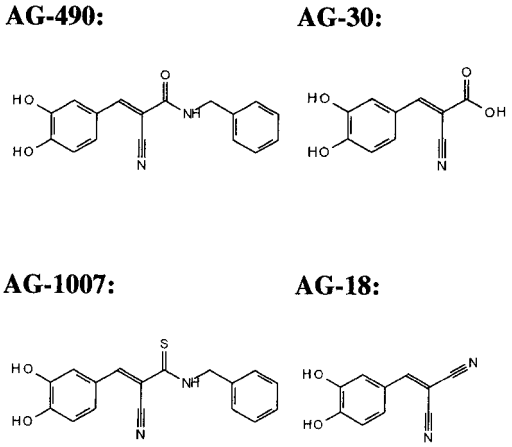


Figure 4. The structures of AG-490 and the other two tyrphostins with negative docking energies, together with the structure of AG-1007.

Results and discussion

Homology modelling

Figure 1 shows that the estimated reliability of our models of Tyk2 and Jak2 is high in large areas of the structures. Most of the nucleotide-binding loop and the activation loop of Tyk2 are predicted with relatively high reliability, but we have some problem areas with large gaps in the sequence alignment (Figure 2). The reliability of the structure prediction of the nucleotide-binding loop and the activation loop of Tyk2 seems to be somewhat higher than for Jak2. The sequence identity between the modelled domains of Tyk2 and Jak2 is 47.9%.

The influence of the choice of template was verified by comparing three different one-template models for both Tyk2 and Jak2. 1QPCA, 1QCFA and 1IR3A were used as templates for Tyk2, while 1BYGA, 1QPCA and 1IR3A were used as templates for Jak2. The other individual templates did not contain sufficient information to construct reasonable homology models when used alone. Structure superpositioning gave a Cα RMSD of 1.31 Å between the three models of Tyk2, and a Cα RMSD of 1.18 Å between the models of Jak2. The Cα RMSD between the five-template models of Tyk2 and Jak2 was 0.75 Å, and increased to 0.77 Å after geometry optimisation. The relatively high RMSD values between the homology models made using one template illustrates the importance of the choice of templates.

Figures 2 and 3 show the multiple alignments of Tyk2 and Jak2, respectively, with their templates. SwissModel makes a structural correction of the alignments. Conformational differences between the templates therefore lead to gaps in the DFG motifs and in the nucleotide-binding loop of Jak2. Figure 2 shows that 1QPCA and 1IR3A, which are both crystal structures of activated kinase domains, give the best alignments with Tyk2 in the activation loop. Hence, Tyk2 is modelled in its active conformation. We see from Figure 3 that 1FPUA and 1QCFA give the best sequence alignments with Jak2 in these regions. Since both 1FPUA and 1QCFA are structures of inactive kinases, Jak2 is modelled in its inactive conformation.

As described by Schindler et al. [13], the selective inhibitor STI-571 binds to the inactive conformation of Abl kinase. We assume that the greater diversity in inactive kinases makes them better targets for selective drug therapy. Attempts to make a model of the inactive conformation of Tyk2 have not yet given reliable results. The model of the active conformation of the Tyk2 tyrosine kinase domain was therefore used.

Docking analysis

As the purpose of the structure modelling of Tyk2 is inhibitor design, the ability of the models to distinguish inhibitors from non-inhibitors is essential. Table 1 shows the results from the computational docking of the 11 tyrphostins in the Jak2-model. Table 1 shows that AG-490 is identified as the most potent inhibitor of Jak2. All the other tyrphostins are predicted to be less potent than AG-490. The separation of AG-1007 from AG-490 is especially interesting, since their structures are very similar. In AG-1007 the carbonyl oxygen of AG-490 is replaced with a sulphur atom. The structures of AG-490 and the other two tyrphostins with negative docking energies are shown in Figure 4, together with the structure of AG-1007. The Model B-factors of the Tyk2 and Jak2 models from SwissModel (see Figure 1) and the WHAT_CHECK report indicate that the model of Tyk2 is at least as good as the model of Jak2. Hence, we assume that the Tyk2-model may be used in inhibitor design.

PASS Analysis

The score plot from the DPLS regression (Figure 5) indicates that the PASSA data is able to predict the correct memberships for the Tyk2- and Jak2 models, and the other kinases. Since 1IR3 and 1QPC are crystal structures of activated kinases, JAK2-1IR3 and JAK2-1QPC lie closer to the Tyk2-models than the other models of Jak2. The models of Tyk2 and Jak2

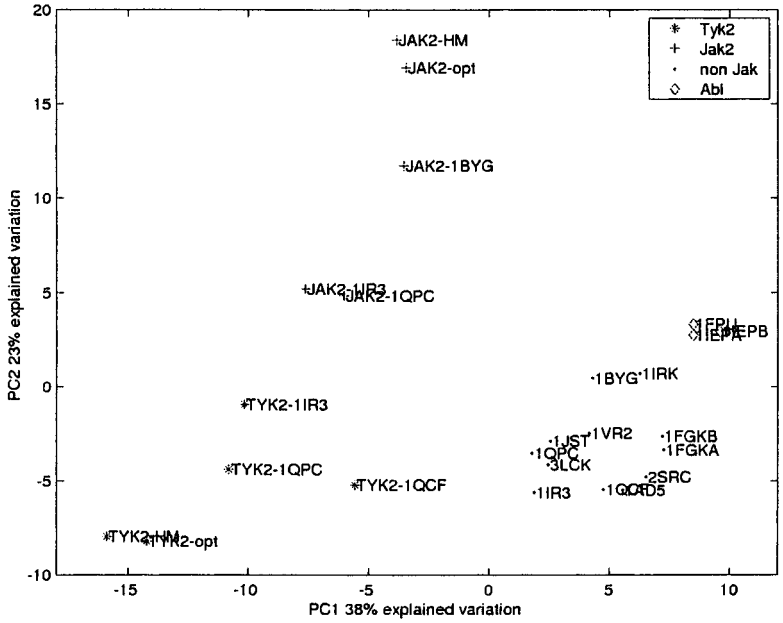


Figure 5. Score plot of the two first principal components (PCs) from the DPLS regression. JAK2-1BYG = Model of Jak2 made using only 1BYGA as template (likewise for the other one-template models). Postfix 'HM' means five-template homology model before geometry optimisation. Postfix 'opt' means the final optimised model made using five templates.

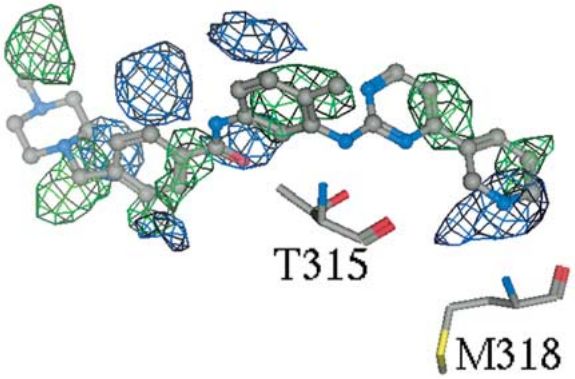


Figure 6. Plot of the regression coefficients for the hydrophilicity (blue) and the hydrophobicity (green) from the DPLS analysis mapped back onto the grid surrounding the ATP binding pocket of Abl kinase in complex with STI-571 (PDB entry 1IEP).

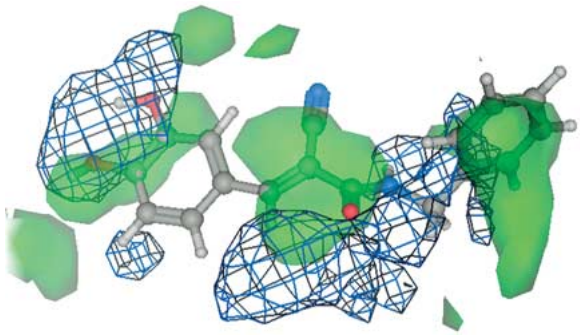


Figure 7. Plot of the regression coefficients for the hydrophilicity (blue) and the hydrophobicity (green) from the DPLS analysis mapped back onto the grid surrounding the model of the ATP binding pocket of Jak2 containing the lowest docking energy conformation of AG-490.

made using five different templates are the most reliable models of these two enzymes. These models are most widely separated from each other and from the other kinases in the score plot. This indicates that these models have properties that are more characteristic to Tyk2 and Jak2, respectively, than the models made using only one template.

The regression coefficients from the DPLS analysis indicate areas where each class of proteins has properties that separate them from the other proteins. Interactions between the protein and inhibitor groups placed in areas of high regression coefficients can contribute to selectivity. We have verified the performance of PASSA by testing whether residues known to interact with selective inhibitors are among the residues identified to be unique. STI-571 is a selective inhibitor of Abl kinase [13,27]. The regression coefficients for Abl were mapped back onto the grid on the X-ray structure of this protein present in PDB entry 1IEP. We see from Figure 6 that the areas of high regression coefficients for the hydrophobicity correspond well to

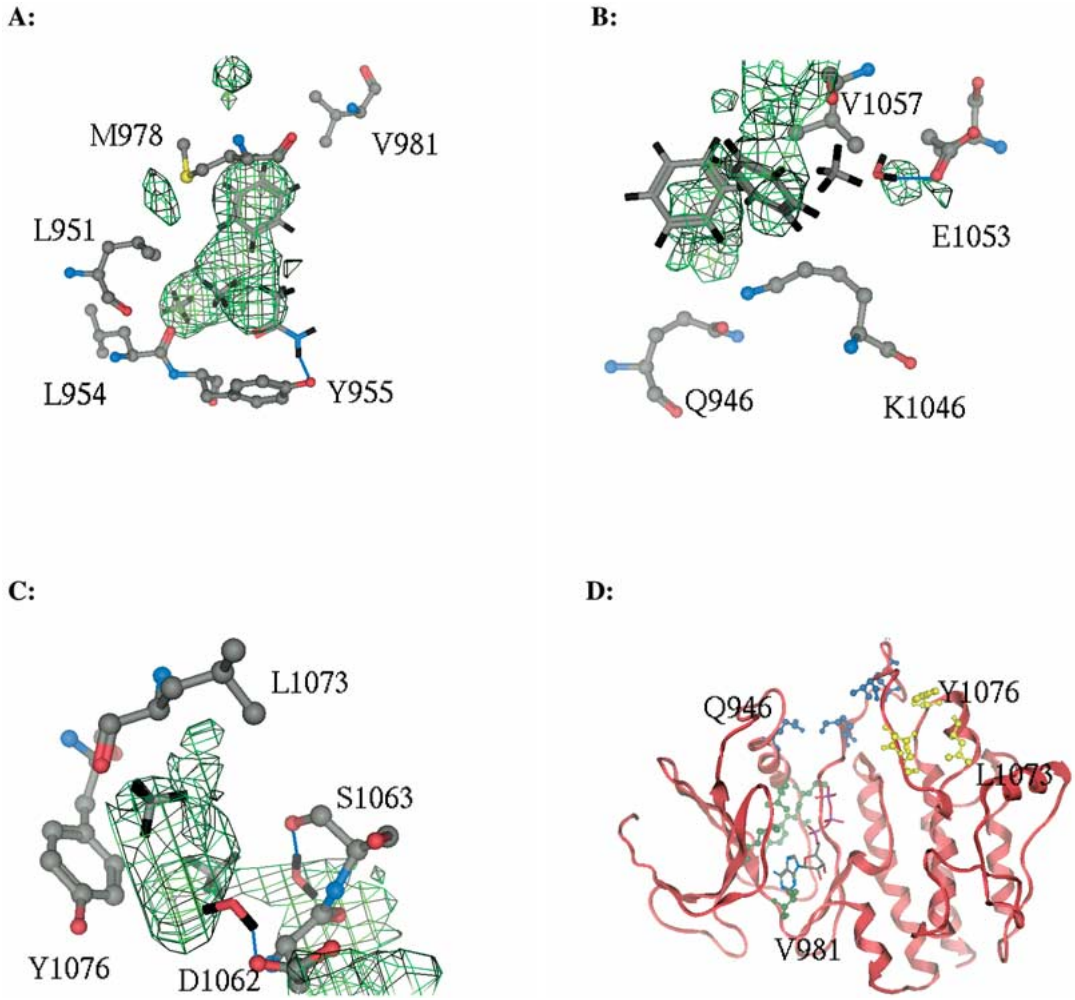


Figure 8. A–C: Plots of the regression coefficients for the hydrophobicity (green) from the DPLS analysis mapped back onto the grid surrounding the model of the ATP binding pocket of Tyk2. Residues identified by PASSA to be unique to Tyk2 are shown, together with selected fragments from the MCSS. Possible hydrogen bonds between MCSS fragments and hydrogen acceptors/donors of Tyk2 identified to be unique are shown as blue lines. D: The residues from Figure 8A (green), 8B (blue) and 8C (yellow) shown together with the result from the computational docking of ATP in the Tyk2 model.

the hydrophobic parts of STI-571. The phenyl-moiety of STI-571 known to interact with T315 in Abl [13] is placed in an area where Abl is particularly hydrophobic compared to the other proteins. The interaction between T315 and STI-571 is known to be important for selectivity [13,53]. According to our results, Abl kinase is particularly hydrophilic in the areas around the carbonyl group of STI-571, and around the nitrogen interacting with M318. A similar analysis was carried out using the homology model of Jak2 and the lowest docking energy conformation of AG-490. AG-490 inhibits Jak2, but none of the other proteins included in this analysis [10,11,28–30]. In Figure 7, the regression coefficients from the DPLS regression are mapped back onto the grid on the model of the ATP binding pocket of Jak2. The results from the PASSA indicate that Jak2 is particularly hydrophilic compared to the other proteins in the areas close to the OH-groups, the NH-group and the carbonyl group of AG-490. We see from Figure 7 that the areas where Jak2 is particularly hydrophobic correspond well to the hydrophobic parts of AG-490. The fact that the interactions between Abl kinase and STI-571, and between Jak2 and AG-490 are identified by PASSA, indicates that this approach may be utilised in the design of selective drugs.

Figure 8 shows the results from the PASS analysis of Tyk2. According to our results, Tyk2 has three

unique hydrophobic pockets that can be utilised by an inhibitor (shown in Figure 8A, B and C, respectively). Similar analysis for the hydrophilicity identified useful hydrogen acceptors and donors close to these pockets. According to our results, interactions with hydrogen acceptors/donors on the following residues can be utilised to achieve selectivity towards Tyk2: Y955, E1053, D1062 and S1063. Fragments from the MCSS placed in regions of high DPLS regression coefficients indicate possible functional groups for a selective Tyk2 inhibitor. These results can be used as a starting point for combinatorial library design, database searching and de novo ligand design. Figure 8A shows that the hydrophobic pocket created by V981, M978, L951, L954 and Y955 can be utilised by an inhibitor having hydrophobic groups pointing towards M978 and L954. In addition, a hydrophobic group with a hydrogen donor or acceptor interacting with the OH-group of Y955 may be advantageous. The hydrophobic pocket shown in Figure 8B can be occupied by a relatively large, aromatic structure, containing a hydrogen donor group in hydrogen-bonding position to E1053. According to our results, a selective inhibitor should also contain a hydrogen donor or acceptor that could interact with the OH-group of S1063, and a hydrogen donor close to D1062. These groups can be connected by e.g. a hydrocarbon chain occupying the space between Y1076 and L1073 (Figure 8C). Figure 8D shows the residues that according to our analysis may be utilised to achieve selectivity towards Tyk2. One can, of course, not guarantee that other protein structures not included in this analysis do not have the same properties as Tyk2 in some of these areas.

In conclusion, we have developed a useful method that identifies binding sites for functional groups that can lead to selectivity. The method has been tested using both X-ray structures and homology models, and appears to be robust against small structural errors caused by the homology modelling process and the computational docking.

Acknowledgements

We want to thank Anders Sundan and Magne Børset at The Department of Cancer Research and Molecular Biology at The Norwegian University of Science and Technology for helpful discussions. We also want to thank The Norwegian Research Council and Amersham Health for financial support.

References

1. Johnson, L.N., Noble, M.E.M. and Owen, D.J., Cell, 85 (1996) 149.
2. Ihle, J.N., Witthuhn, B.A., Quelle, F.W., Yamamoto, K. and Silvennoinen, O., Annu. Rev. Immunol., 13 (1995) 369.
3. Pellegrini, S. and Dusanter-Fourt, I., Eur. J. Biochem., 48 (1997) 615.
4. Van der Geer, P., Hunter, T. and Lindberg, R.A., Annu. Rev. Cell Biol., 10 (1994) 251.
5. Richter, M.F., Dumenil, G., Uze, G., Fellous, M. and Pellegrini, S., J. Biol. Chem., 273 (1998) 24723.
6. Hubbard, S.R. and Till, J.H., Annu. Rev. Biochem., 69 (2000) 373.
7. Pautsch, A., Zoephel, A., Ahorn, H., Spevak, W., Hauptmann, R. and Nar, H., Structure, *9* (2001) 955.
8. Heldin, C.H., Cell, 80 (1995) 213.
9. Anderson, K., Semin. Oncol., 26 (1999) 10.
10. Meydan, N., Grunberger, T., Dadi, H., Shahar, M., Arpaia, E., Lapidot, Z., Leeder, J.S., Freedman, M., Cohen, A., Gazit, A., Levitzki, A. and Roifman, C.M., Nature, 379 (1996) 645.
11. Wang, L.H., Kirken, R.A., Erwin, R.A., Yu, C.R. and Farrar, W.L., J. Immunol., 162 (1999) 3897.
12. Lindauer, K., Loerting, T., Liedl, K.R. and Kroemer, R.T., Protein Eng., 14 (2001) 27.
13. Schindler, T., Bornmann, W., Pellicena, P., Miller, W.T., Clarkson, B. and Kuriyan, J., Science, 289 (2000) 1938.
14. The RCSB Protein Data Bank, http://www.rcsb.org/pdb/.
15. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235.
16. Bajorath, J., Stenkamp, R. and Aruffo, A., Protein Sci., 2 (1993) 1798.
17. Chothia, C. and Lesk, A.M., EMBO J., 5 (1986) 823.
18. Lesk, A.M. and Chothia, C., J. Mol. Biol., 136 (1980) 225.
19. Chothia, C. and Lesk, A.M., J. Mol. Biol., 160 (1982) 309.
20. Read, R.J., Brayer, G.D., Jurásek, L. and James, M.N.G., Biochemistry, 23 (1984) 6570.
21. Pastor, M. and Cruciani, G., J. Med. Chem., 38 (1995) 4637.
22. Miranker, A. and Karplus, M., Proteins Struct. Func. Genet., 11 (1991) 29.
23. Edelsbrunner, H., Facello, M., Fu, R. and Liang, J., Proceedings of the 28th Hawaii International Conference on Systems Science¸ Maui, January 1995, pp. 256-264.
24. Liang, J., Edelsbrunner, H. and Woodward, C., Protein Sci., 7 (1998) 1884.
25. Martens, H. and Martens, M. Multivariate Analysis of Quality, An Introduction, John Wiley & Sons, Ltd., Chichester, 2000.
26. Klebe, G., Abraham, U. and Mietzner, T., J. Med. Chem., 37 (1994) 4130.
27. Zimmermann, J., Buchdunger, E., Mett, H., Meyer, T. and Lydon, N.B., Bioorg. Med. Chem. Lett., 7 (1997) 187.
28. Xuan, Y.T., Guo, Y.R., Han, H., Zhu, Y.Q. and Bolli, R., Proc. Natl. Acad. Sci. USA, 98 (2001) 9050.
29. Kirken, R.A., Erwin, R.A., Taub, D., Murphy, W.J., Behbod, F., Wang, L.H., Pericle, F. and Farrar, W.L., J. Leukocyte Biol., 65 (1999) 891.
30. Bright, J.J., Du, C.G. and Sriram, S., J. Immunol., 162 (1999) 6255.
31. Guex, N. and Peitsch, M.C., Electrophoresis, 18 (1997) 2714.
32. Peitsch, M. C., Biochem. Soc.Trans., 24 (1996) 274.
33. Guex, N., Diemand, A. and Peitsch, M.C., TiBS, 24 (1999) 364.
34. Peitsch, M.C., Bio/Technology, 13 (1995) 658.

35. Molecular Operating Environment™, Version 2001.01, Chemical Computing Group, Inc., 2001.
36. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.
37. Charifson, P.S. Practical Application of Computer-Aided Drug Design, Marcel Dekker, Inc., New York, 1997.
38. Vriend, G., J. Mol. Graph., 8 (1990) 52.
39. Hooft, R.W.W., Vriend, G., Sander, C. and Abola, E.E., Nature, 381 (1996) 272.
40. Swiss-PdbViewer, Version 3.7b2, Glaxo Wellcome Experimental Research, 2001.
41. Halgren, T.A., J. Comp. Chem., 17 (1996) 490.
42. Still, W.C., Tempczyk, A., Hawley, R.C. and Hendrickson, T., J. Am. Chem. Soc., 112 (1990) 6127.
43. Qiu, D., Shenkin, S., Hollinger, F.P. and Still, W.C., J. Phys. Chem., 101 (1997) 3005.
44. Schaefer, M. and Karplus, M. A., J. Phys. Chem., 100 (1996) 1578.
45. Baxter, C.A., Murray, C.W., Clark, D.E., Westhead, D.R. and Eldridge, M.D., Proteins Struct. Funct. Genet., 33 (1998) 367.
46. Halgren T.A., J. Comp. Chem., 20 (1999) 730.
47. Needleman, S. B. and Wunsch, C. D., J. Mol. Biol., 48 (1970) 443.
48. Gonnet, G.H., Cohen, M.A. and Benner, S.A., Science, 256 (1992) 1433.
49. Shapiro, A., Botha, J.D., Pastore, A. and Lesk, A.M., Acta Cryst., A48 (1992) 11.
50. Labute, P. and Williams, C., J. Med. Chem., 44 (2001) 1483.
51. MATLAB™, Version 5.3, MathWorks, Inc., 1999.
52. PLS Toolbox®, Version 2.1, Eigenvector Research, Inc., 2001.
53. Gorre, M.E., Mohammed, M., Ellwood, K., Hsu, N., Paquette, R., Rao, P.N. and Sawyers, C.L., Science, 293 (2001) 876.

Paper III

Homology-Based Modelling of Targets for Rational Drug Design

Heather Wieman[1], Kristin Tøndel[2], Endre Anderssen[2] and Finn Drabløs[1*]

[1]Department of Cancer Research and Molecular Medicine, Faculty of Medicine, MTFS, Norwegian University of Science and Technology, N-7489 Trondheim, Norway, [2]Department of Chemistry, Norwegian University of Science and Technology, Sem Sælands v. 14, N-7491 Trondheim, Norway

Abstract: The current status in rational drug design using homology-based models is discussed, with focus on template selection, model building, model verification and strategies for drug design based on model structures. A novel approach for identification of unique binding site features from homology-based models, Protein Alpha Shape Similarity Analysis (PASSA) is described.

Keywords: Homology Model; Drug Design; Template Selection; Model Verification; Protein Alpha Shape Similarity Analysis

INTRODUCTION

Rational drug design is an important concept in pharmaceutical research. The goal is to identify a key drug target based on a thorough understanding of regulatory networks and metabolic pathways, and to design a highly specific drug based on the known three-dimensional (3D) structure of that target. The flood of data from large-scale genome oriented projects is bringing this concept closer to reality. The detailed mapping of genome sequences, regulatory networks and metabolic pathways combined with single nucleotide polymorphism (SNP) data, biological samples or health records makes it easier to identify optimal drug targets. Access to high-quality 3D structures of these targets is a good starting point for rational design of novel drugs.

There are several examples of rational drug design using targets with known 3D structure, including the HIV protease inhibitors amprenavir (Agenerase) and nelfinavir (Viracept) [1-3] and the influenza virus inhibitor zanaminivir (Relenza) [4]. Structure-based drug design has also been applied for example in the design of inhibitors of protein kinases [5] such as Abl kinase [6], CDKs [7], EGFR kinase [8], Lck [9] and Src [10].

X-ray crystallography is the main method for structure determination of proteins. This can be a time-consuming process, and it will succeed only if it is possible to find suitable conditions for growing crystals. This can therefore easily become a bottleneck in drug design projects. However, structural domains of proteins can be classified into classes of similar folds, and the number of protein folds actually used by nature seems to be limited [11]. Experimental structure data have been generated for a large fraction of these possible folds, and ongoing structure determination efforts focus on making this mapping as complete as possible. This makes homology-based modelling of protein structures a realistic and relevant alternative to experimental structure determination.

*Address correspondence to this author at the Finn Drabløs, Department of Cancer Research and Molecular Medicine, Faculty of Medicine, MTFS, Norwegian University of Science and Technology, N-7489 Trondheim, Norway; Tel: +47 73 59 88 42; Fax: +47 73 59 88 01; E-mail: finn.drablos@medisin.ntnu.no

Comparative modelling is often used as a neutral alternative to *homology modelling*, which implies an evolutionary relationship between target and templates. Homology modelling has been used successfully in several drug design projects. Enyedy *et al.* [12] have utilised a homology model of Bcl-2 to identify a novel class of inhibitors by structure-based computer screening. Furet *et al.* [13] successfully applied homology-based modelling for rational design of inhibitors of Cyclin-dependent kinase 1 (CDK1). A modelled structure of an antagonist-bound retinoic acid receptor based on the structure of estrogen receptor has been applied for virtual ligand screening, resulting in the discovery of three novel ligand candidates [14], and homology modelling of Falcipain-2 provided information that led to the discovery of new drug leads against malaria [15].

It is a matter of discussion whether homology models are accurate enough to be utilised in ligand screening and design. It is at least important to use methods that are robust against small structural errors. Recently, Schafferhans and Klebe [16] published a method for computational docking of ligands into protein binding sites that is especially suited for protein structures derived by homology modelling. This method incorporates ligand information into the protein structure modelling procedure. Another drug design method, PASSA, has also been developed specifically for use on homology models. This method uses several alternative homology models for the same protein together with structures of other, related proteins to single out unique features of the target protein [17]. However, such approaches do not decrease the importance of high quality models of potential targets.

HOMOLOGY MODELLING

Homology modelling is based on the observation that the 3D structure of homologous proteins is more conserved than sequence [18]. Chothia and Lesk [19] investigated the relation between sequence conservation and structural similarity for 32 pairs of homologous proteins, and concluded that a protein structure can provide a close general model for other proteins if the sequence similarity is greater than 50%. When the sequence identity drops to 20%, there will be large structural differences. However, the active sites

1389-5575/04 $45.00+.00

can have very similar geometries, even for distantly related proteins [20,21].

The methodology itself can be described in four steps (illustrated in Fig. (**1**)): Identifying a suitable template, making an optimal target-template alignment, building the model and validating the model. Protein structure prediction and homology modelling has recently been reviewed by Schonbrun *et al.* [22] and Al-Lazikani *et al.* [23].

Template Identification

The first step is matching the protein sequence of interest (the target) to experimentally determined structures, in order to find at least one protein (the template) for which we can assume that it has the same 3D structure as the target. This is normally based on sequence similarity. Heuristic search methods such as BLAST [24] and FASTA [25] are often used in the initial template-finding step, because these methods are fast and well tested. In difficult cases more sensitive fold recognition methods, which utilise techniques such as Hidden Markov Methods, Neural Networks, iterated searches (e.g. PSI-BLAST [26]), and evolutionary information can be used to scan a structural database for suitable templates [27]. In particular when no close homologues can be found, the increased sensitivity from these methods may allow more potential templates to be identified. This may improve the general reliability of the model, and it may help in identifying structurally conserved regions. For the same reason it is generally advantageous to use several fold recognition methods in parallel, as alternative algorithms may retrieve slightly different data sets and alignments [28].

Alignment

After identification of the best templates for modelling, an optimal alignment must be made. This seems to be the most crucial step in homology modelling [22,29]. Here "optimal" means that corresponding sequence positions in target and template are identified, so that the predicted structure of the target, based on the template, is as similar as possible to an experimental structure of the same target. Identification of corresponding sequence positions in terms of evolution will at least give a close approximation to an optimal alignment. It is important to realise that the sequence alignment of target with respect to a template identified by a search method or fold recognition method may be sub-optimal with respect to modelling. Different score matrices are needed in order to get optimal alignments for homology modelling as compared to fold recognition [30], possibly because fold recognition needs to focus on conserved regions whereas homology modelling needs to take all regions into account. Hence, alignments generated from fold recognition methods often require refinement in order to be utilised for modelling.

The Smith-Waterman algorithm uses dynamic programming to find an optimal alignment between two sequences, given a scoring matrix and a gap model [31,32].

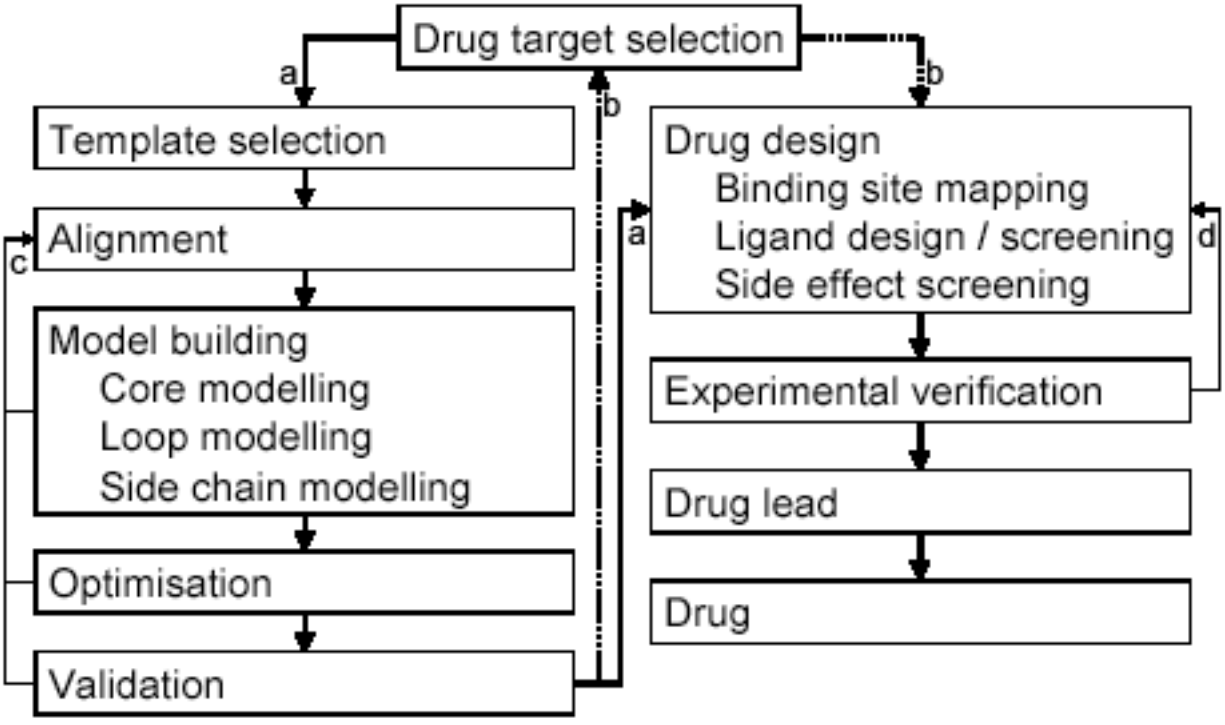


Fig. (1). Key steps in the homology modelling and drug design pathway.

The modelling and design process normally starts with a suitable drug target, the 3D structure of the target is predicted through homology modelling and the model is used for rational drug design (a). However, it is also possible to start with general high-throughput modelling using all potential targets, followed by target selection based on these models (b). In both sections (homology modelling and drug design) there are feedback loops, where e.g. model quality may be used to improve alignment (c) or experimental data may have influence on the drug design strategy and lead optimisation (d).

However, the scoring matrix and gap model represents a simplified model of evolution, and a mathematically optimal solution may still be wrong from an evolutionary perspective. The use of alignments based on multiple sequences is recommended, as this will highlight evolutionary relationships, and increase the probability that corresponding sequence positions are correctly aligned. Moreover, evolutionary information incorporated into sequence profiles greatly increases the alignment accuracy, bringing the alignment closer to the 'true' structural alignment [33]. ClustalX [34], Poa [35], Dialign [36,37] and T-Coffee [38] are important multiple alignment tools. It has been reported that for cases of low sequence identity, Dialign produces the most accurate alignments, whereas T-Coffee is more robust in cases of higher sequence identities [39]. Improved performance can be achieved by combining several alignment strategies [40,41]. Other interesting methods include machine learning [42], fast Fourier transform [43] and improved score matrices built from structural superpostion data [44]. New scoring functions have also been developed to give a quantitative measure of alignment accuracy [45].

Using structurally aligned templates as a starting point for the multiple sequence alignment will improve the alignment quality if sequence similarity is low [23,40,46]. Alignment programs such as DALI [47], STRUCTAL [48] and LOCK [49] are examples of structural alignment methods for aligning multiple templates.

Regardless of which program is used, the quality of the alignment should always be verified. However, this is closely related to verification of the homology model itself, and will therefore be discussed there.

Model Building

Model building consists of three main steps. The homology is important mainly when building the core of the protein. Loop modelling is basically *de novo* model building, whereas side chain (re)modelling mainly is an optimisation step. Reliable identification of structurally conserved core regions versus variable loop regions is an important aspect of this process [50].

There are currently three important approaches for building the core region from alignments. Rigid body superposition constructs the model from a few core sections defined by the average of Cα atoms in the conserved regions. Distance geometry uses spatial restraints obtained from the alignment. Segment matching uses a database of short segments of protein structure, with energy or geometry rules, or some combination. It has been shown that when used optimally, accuracies are similar for most modelling methods [51]. Some commonly used programs for homology modelling are listed in Table 1.

SwissModel [52,63] is a popular implementation of the rigid body approach. ProModII [64] generates a model framework based on the topological arrangement of corresponding atoms to the given templates. The backbone is rebuilt based on the positions of Cα atoms, using a library of backbone elements derived from high quality X-ray structures. Incomplete loops and incomplete or missing side chains are rebuilt before the models are energy minimised with molecular mechanics (MM).

Homology modelling in MODELLER [56] is based on satisfaction of spatial restraints. Distance and dihedral angle restraints on the target structure are generated, based on the alignment to the template structure. Corresponding distances and angles between aligned residues in the template and the target structures are assumed to be similar. Restraints on bond lengths, bond angles, dihedral angles and nonbonded atom-atom contacts are also derived from statistical analysis of the relationships between Cα atoms, solvent accessibilities and side-chain torsion angles in known protein structures. The restraints are expressed as probability density functions (pdfs). These pdfs are combined to give a molecular function, which is optimised using a combination of energy minimisation with molecular dynamics and simulated annealing.

Table 1. Some Commonly Used Homology Modelling Programs

Method	Type[a]	Ref	Url
SWISS-MODEL	RBS	[52]	http://www.expasy.org/swissmod/SWISS-MODEL.html
WHATIF	RBS	[53]	http://www.cmbi.kun.nl/whatif/
COMPOSER	RBS	[54-58]	http://www.tripos.com/
CONGEN	RBS	[59]	http://www.congenomics.com/
InsightII/Homology	RBS	[60]	http://www.accelrys.com/
TURBO-FRODO	RBS		http://afmb.cnrs-mrs.fr/TURBO_FRODO/
JACKAL	RBS		http://trantor.bioc.columbia.edu/~xiang/jackal
ICM-Homology	RBS	[61]	http://www.molsoft.com/
Look/GeneMine	SM	[62]	http://www.bioinformatics.ucla.edu/genemine/
MODELLER	SR	[56]	http://www.salilab.org/modeller/modeller.html
InsightII/Modeler	SR	[56]	http://www.accelrys.com/

[a] RBS – Rigid body superposition, SM – Segment matching, SR – Spatial restraints

The LOOK software package [62] uses Segment Match Modeling (SegMod) to generate homology models by fragment based assembly [65]. SegMod uses a powerful fragment-matching algorithm to find the appropriate structural segments derived from known 3D structures. It utilised both backbone and side chain information from the fragments to obtain a complete model. After building 10 individual models, the averaged model is then minimised using molecular mechanics. SegMod handles insertions and deletions during model building by searching for compatible fragments.

Separate steps are often used for predicting loops (loop libraries or *ab initio* loop building) [66-69], and modelling side chains [68,70-74]. These methods can be used in combination with any of the core modelling techniques.

Although functionally important regions usually are well conserved, flexible loop regions may often contribute significantly in defining specificity. Accurate loop modelling may therefore be important for the usefulness of the homology model. However, existing methods are generally not reliable for loops longer than 5 residues [75]. Loops are often too short to provide sufficient information about their local fold, and segments of up to 9 residues sometimes have entirely unrelated conformations in different proteins [76,77]. Identification of optimal anchor groups seems to be an important step in loop prediction [78,79]. *Ab initio* loop prediction has recently been discussed by Galaktionov *et al*. [69].

After the initial model building, the model can be optimised with molecular mechanics software using either energy minimisation or molecular dynamics methods, or a combination. However, optimisation methods will in general not bring models closer to the true structure [22]. In fact, with extensive refinement homology models actually tend to get worse [80]. Recent data from Flohil *et al*. indicate that some improvement may be gained if long time scale simulation with explicit inclusion of water molecules is used [81]. However, since the roles of optimisation procedures in improving structural quality are still debated [51,68,70], they should be used with caution. Particular care has to be taken when domains, rather than full structures, are modelled. In domain modelling, the positions of any atoms forming an interface to a missing domain should be fixed during energy minimisation. Free movement in these regions can lead to side chain conformations that are preferable energetically, but not possible in the real protein structure because of interactions with residues in the missing domain.

It is still relatively unclear which approach generates the best model. Since 1994 several modelling groups have participated in a bi-annual evaluation project, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [82]. The groups model proteins that are in the process of being solved experimentally, but not yet have been released for publication. The submitted models are later compared to the then released structures to determine which modelling methods have been most successful. There is also a web-server, EVA-CM (http://www.pdg.cnb.uam.es/eva/cm/) which is designed to evaluate protein structure prediction and modelling servers in 'real time' [83,84]. This server evaluates the 'black box' modelling programs. These programs often limit the number of templates used and impose limitations on manual intervention.

Validation of Models

The accuracy of a model depends upon the sequence similarity it shares with the template. Models with >50% sequence identity to templates are normally of high quality, with ~1 Å root mean square (RMS) error for main chain atoms (equal to medium-resolution NMR or low resolution x-ray structures). Models that have 30 – 50% sequence identity are normally of medium accuracy with an RMS of ~1.5 Å [51,76,85]. Typical errors include problems with side-chain packing, core distortion, loops, and misalignment.

Several validation checks are used for assessing model quality. The most common checks pertain to geometric and stereochemical measurements: covalent geometry (bond lengths and angles), planarity, chirality, phi/psi preferences, chi angles, non-bonded contact distances, unsatisfied donors/acceptors etc [86,87]. Ramachandran plots can provide an overall view of phi/psi values and is a good indicator of the global quality of the model [88]. Quality checks such as these are present in standard crystallographic and NMR software packages as well as in software designed for molecular modelling (e.g. WHATIF and PROCHECK) [53,89]. However, this analysis only indicates the presence of unusual conformations in the structure. Even an incorrect alignment may end up with very reasonable local geometry. Hence, additional tests are needed, in particular for models based on templates with low sequence similarity, where the possibility for misalignment is significant. This is a quite general problem, an interesting example of a misalignment error was recently identified in an experimental 3D structure [90].

Many of these tests are basically fold recognition methods scoring the compatibility between the target sequence and the predicted 3D structure. Sippl *et al.* uses an inverse Boltzmann principle to calculate a mean force potential by 'threading' the target sequence onto structures [91], measuring how well the primary sequence fits the given three-dimensional structure. A related approach tests model correctness by way of a 3-D profile [92]. The 3-D profile of the structure describes the structural environment of each residue. This can be used to score compatibility of any amino acid sequence with that structure. Yet another quality assessment algorithm takes into consideration geometrical parameters of a given structure and then calculates the local, buried and contact energy via statistical potentials of mean force [93-96]. This method has been used in homology modelling to evaluate alternative protein models based on different alignments and as a detector of problematic regions within the protein structure. Another validation measure, designed directly from the results of CASP3, seeks to find the largest subset of Cα atoms of the model that can be superpositioned well with the template structures it was modelled from. The normalised score reflects a rough quality measure of the model [97].

The validation checks have to be viewed in light of the validation of the template structure itself. Crystallographic structures are also prone to error, and whatever discrepancies

introduced through the chemical structure determination will most likely also arise in any model based on that structure. The best approach is to gather as much information from as many sources as possible, for both model and templates.

DRUG DESIGN

Given a suitable model of the 3D structure of a potential target, the drug design step tries to find the optimal compound for moderating the normal function of the target in a selective and normally reversible way. In addition to this, several physical criteria have to be met, related to production, uptake, degradation etc. Here we will focus on the actual ligand design, in particular on methods that may improve selectivity. In order to design a ligand for a given target possible interaction sites for ligands have to be identified and the properties of these sites have to be mapped. However, considering only the target protein may be a mistake. Many drugs have recently been withdrawn from late stage testing due to off target effects [98]. Hence, to achieve selectivity and avoid side effects, knowledge of related binding sites is also important. Homology modelling makes this practical, as dozens or hundreds of protein structures can be obtained. If such massive amounts of structural data are to be useful, data analytical methods are needed that aid the interpretation of structural data.

Mapping of Binding Sites

Numerous methods for mapping protein binding sites exist, the majority of which utilise calculations of interaction energies between the protein and small, molecular probes. Binding site analysis is a prerequisite for effective database searches, docking, and *de novo* ligand design. The field of binding site analysis has recently been reviewed [99]. Therefore, this review will focus on mapping strategies that enable comparison of numerous structures for the purpose of understanding selectivity, in particular Multiple Copies Simultaneous Search (MCSS), GRID and Protein Alpha Shape Similarity Analysis (PASSA).

Multiple copies simultaneous search is a method for finding favourable interaction sites in a protein cavity [100]. The idea behind MCSS is to place a large number of copies of one or more probe molecules into the active site of the target. These probes are placed randomly around the active site atoms and are assumed not to interact with each other (Fig. (**2**), left). Next, a special energy minimisation protocol is used to refine the initial placement. The receptor atoms may be kept fixed, or be subject to the average forces of the probes [101]. Each probe is subject to the full force of the receptor but not forces from the other probes. Once stable receptor and fragment geometries have been found, fragments with high energies are deleted. The resulting low energy fragments and how they interact with the receptor can then be analysed (Fig. (**2**), right). The probe molecules are free to move and will have migrated towards regions of favourable interaction with the receptor. This identifies regions of strong interactions that may be used by a ligand. It also gives information on favourable orientation of functional groups. This is useful for *de novo* ligand design as the low energy fragments can be used as starting points. However, a more systematic and complete mapping of the binding site may be necessary, since this random search strategy may not find all relevant interactions.

One of the most common methods for mapping ligand binding sites in proteins is GRID [102], which uses a regular grid spanning the binding site. At each grid point the interaction energy between the protein and a probe group placed on the grid point is computed using a molecular mechanics energy function. Parameters for probes representing various functional groups have been developed [103,104]. The results can be visualised as contour plots of the interaction energies for different probes, and highly detailed potential maps of binding sites may be produced. The low energy contours indicate where functional groups of a ligand are likely to be placed. GRID has been used to suggest functionality for both antibacterial [105,106] and antiviral drugs [107,108]. Both GRID and MCSS have been compared to experimental binding of small molecules by crystallising the same protein in various solvents [109]. It was found that both methods identified approximately the same interaction site, but most results were not reproduced experimentally. In some cases MCSS predicted the correct orientation of the probe, but the predicted orientation of large hydrophobic probes was often wrong. The major reason for the discrepancies between experimental and

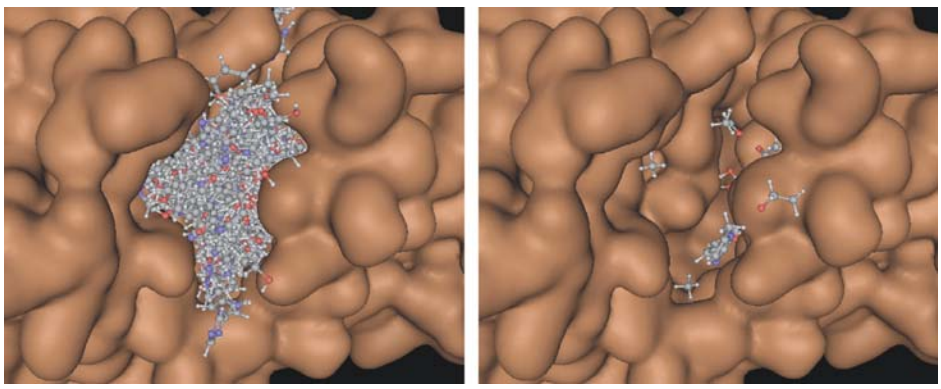


Fig. (2). MCSS.

MCSS mapping of a protein cavity. A large number of different small molecules are placed inside the binding cavity of the protein (left frame). A special energy minimisation procedure is run (see main text). Once the energy minimum has been found, the positions and orientations of low energy molecules may be inspected (right frame).

computational results is believed to be neglect of solvent and entropy effects in the computational models.

Compared to MCSS, the GRID method has the disadvantage that the fragments are not free to move away from their grid point to a more optimal location. However, the fixed grid has a major advantage with respect to comparability. Many related proteins can be superpositioned and the same grid used for all of them. Interesting differences between related binding sites can be identified by comparing energy maps. Thus, using GRID on multiple proteins can aid the development of ligands selective for a particular protein target. The data from the GRID computations can be analysed by Principal Component Analysis (PCA) to find the most important structural differences to take into consideration for the design of a selective inhibitor. The data analysis tools used to analyse GRID results have been refined by changing the weighting of data from different probes. This has been applied to the design of selective inhibitors of both serine proteases [110] and metalloproteinases [111]. Additional work using homology models has been done on human cytochromes [112].

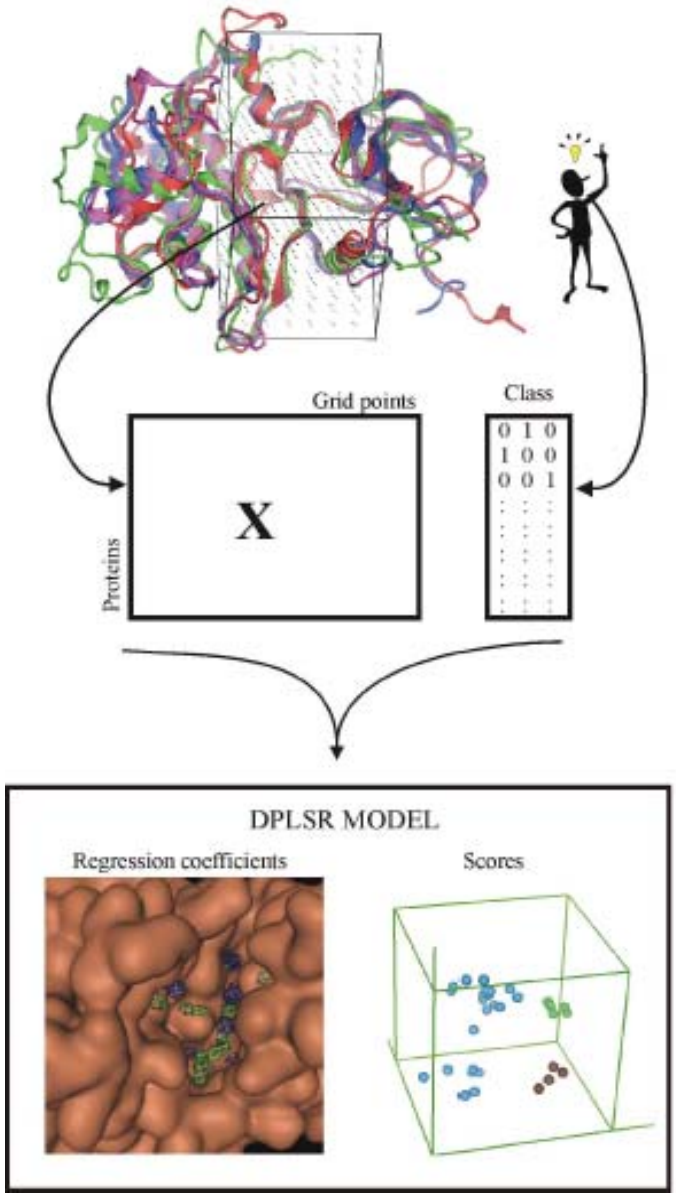


Fig. (3). PASSA.

Protein structures are aligned to maximise the overlap in the active site. A regular grid is placed surrounding the active sites and the alpha shape density of each protein is computed at each grid point. The density data form the matrix **X**. The user also specifies a number of classes, and assigns each protein to a class. The alpha densities and the class data are analysed by DPLSR and a model is produced. Interpretation of the model consists of two parts: Regression coefficients and scores. Mapping the regression coefficients back on a protein structure may indicate which regions may contribute to selectivity. Regions may be colour coded by their lipophilic or hydrophilic nature. The scores provide an alternative picture of the model. In the scores space, every protein is represented by one point. Visualising the distribution of proteins in the scores space is useful for discovering clusters, highly deviant structures and to evaluate the structural diversity in a set of proteins.

The use of homology models as the basis for GRID calculations requires some special considerations, specifically if multiple models are to be used in the design of selective ligands. Computing the interaction energy requires precisely defined atomic charges for all atoms, protonation states, and correct placement of hydrogen atoms. The very steep gradients of most force fields close to nuclei may cause instabilities in the PCA models and inflate the effects of small errors in the homology models or the superpositioning.

Protein Alpha Shape Similarity Analysis (PASSA) is an alternative to GRID, developed particularly for use with homology models in the design of selective ligands. This method uses geometrical objects known as alpha spheres to construct a representation of the active site. An alpha sphere is a sphere that contacts four atoms on its surface and has no atoms contained internally. Small alpha spheres correspond to densely packed regions in the protein, while very large spheres are found on the protein surface. In the typical binding pocket however, medium sized spheres are found. Clusters of medium sized spheres will thus correspond to the binding cavities of the protein. Alpha spheres have proven useful for identifying the binding pockets in a number of proteins, and the centres of alpha spheres have been found to correspond well with the placement of atoms in bound ligands [113]. Alpha shapes are determined geometrically, using only the positions and radii of the heavy atoms. This eliminates the need for placing hydrogens and determining protonation states and partial charges. The alpha spheres are classified as hydrophobic or hydrophilic depending on the protein atoms they contact.

PASSA converts the discreet information contained in the placement of alpha sphere centres and protein atoms to a continuos field using a gaussian density estimate. ”Dummy” atoms placed at each alpha sphere centre are assigned weights for either the hydrophobic or the hydrophilic field, according to the alpha sphere class. The use of gaussian functions with a very simple partitioning according to the hydrophilic or hydrophobic nature of the alpha spheres reduces some of the problems associated with traditional force field models. Gaussian functions have neither steep derivatives nor singularities. The less detailed representation may also be more robust against the errors typically present in homology models. Analysis of data from gaussian fields typically produce contour plots that are less fragmented and easier to interpret than those produced using force field models [114].

PASSA has been used to suggest properties of a selective inhibitor of Tyrosine kinase 2 (TYK2) and also to understand the basis of the selectivity of STI571, a selective Abl kinase inhibitor [17]. In this work, Discriminant Partial Least Squares Regression (DPLSR), rather than PCA, is used to analyse the field data (Fig. (**3**)). DPLSR enables the user to guide the analysis towards features relevant for selectivity towards a specific protein or group of proteins. This is done by dividing the protein structures included in the analysis into classes, typically a ‘target’ class, containing the structures one wishes to develop a ligand for, and an ‘other’ class. The ‘other’ class contains proteins related to the target, but for which a low affinity is desired. Any class scheme may be used e.g. in exploratory work looking for a suitable drug design target. In some cases, a single protein structure may even belong to more than one class. When analysing homology models in this manner, it is advantageous to use more than one model of each protein, particularly if several templates of comparable sequence identity are available. If several independent structures exist in both the ‘target’ and ‘other’ classes, cross validation of the DPLSR model can be used to assess the stability of the model parameters. Thus, the influence of errors in the homology modelling may be gauged. DPLSR works by extracting a low dimensional subspace from the PASSA data that can explain the class structure. Typically relatively few dimensions are needed to separate the classes. This enables visualisation of the relationship between the structure models and easy discovery of clusters or deviant structures. DPLSR models can represent the differences between the protein(s) of interest and all other proteins in the study as a single vector of beta coefficients. The beta coefficients can be visualised as contours in the original 3D space of the protein structures. Spatial regions that may form the basis of selectivity may thus be identified. When designing a TYK2 inhibitor, PASSA was used in combination with MCSS. The plots of the regression coefficients from PASSA were used to guide the selection of MCSS fragments towards those fragments that may contribute to selectivity as well as affinity. This use of combined knowledge of affinity and selectivity is a good starting point for both database searches and *de novo* ligand design, simplifying the task of designing a selective inhibitor.

Database Screening

Once possible interaction sites for a selective inhibitor have been identified, databases of already existing drugs can be searched in order to find a drug molecule that fits the receptor binding site [115]. A number of such databases exist, such as The Cambridge Structural Database [116], the database of The National Cancer Institute (http://cactus.nci.nih.gov/), the Available Chemicals Directory (MDL Information Systems) and PDBsum (which includes a database of ligands from the RCSB Protein Database) (http://www.biochem.ucl.ac.uk/bsm/pdbsum/). The hits from the database searching can then be evaluated further by molecular docking. Available docking programs include AutoDock [117], DOCK [118], FlexX [119], GOLD [120], LUDI [121] and MOE-Dock (Chemical Computing Group Inc). A version of FlexX suited for combinatorial library docking, FlexX^c, has also been developed [122]. Recently, new docking methods especially suited for use with homology modelled protein structures have been developed. Schafferhans and Klebe [16] use gaussian functions to represent the physico-chemical properties of the receptor and the ligand, and optimise the overlap between the functional description of the receptor binding site and the ligand. Another docking method that utilises gaussian functions is the method developed by McGann *et al*. [123], that acts as a filter to reduce the search space for other docking methods. This method only accounts for shape, and minimises steric clashes between the receptor and ligand atoms. The method developed by Wojciechowski and Skolnick [124] uses a discretisation of the structural models together with an averaging of the structural details and a smoothing of the potential energy surface to compensate for structural errors. Both steric and chemical complementarity

between the ligand and the receptor is sought using a grid-based search. A complete cover of existing docking and virtual screening methods is outside the scope of this review, but the topic has recently been reviewed e.g. by Taylor *et al.* [125], Lyne [126] and Bajorath [127].

To limit the computational time, docking simulations have traditionally been carried out with a fixed protein structure. When using protein structure models built by homology modelling, it is especially important to allow for protein flexibility, since this can reduce the impact of small structural errors. Homology models are built using X-ray structures of other proteins as templates. These are often co-crystallised with a ligand, which induces ligand-specific conformational changes in the protein. Using a rigid protein structure might thus prevent us from identifying optimal binding modes for alternative ligands. Some methods, such as the method developed by Leach [128] and the "Mining Minima Optimizer" method developed by Kairys and Gilson [129] use side-chain flexibility. Anderson *et al.* [130] developed an algorithm for identifying regions where conformational adaptation to a ligand is likely to occur. During the docking simulations the side-chains of these residues are allowed to move.

Recently, some new methods have been developed, that take protein backbone flexibility into account. The majority of these methods utilise multiple protein structure models in the calculations. Österberg *et al.* [131] incorporated protein flexibility and structural water heterogeneity into the docking simulations using an ensemble of protein structures. In the "Relaxed Complex Method", developed by Lin *et al.* [132,133] a long molecular dynamics (MD) simulation of the unliganded receptor is carried out, followed by a rapid docking of candidate ligands to a large ensemble of the receptor's MD conformations. The FlexE approach [134] is based on a united protein description generated from an ensemble of protein structures. For varying parts of the protein, discrete alternative conformations are explicitly taken into account, which can be combinatorially joined to create new protein structures. Broughton combined the use of statistical analysis of conformational samples from short-run protein molecular dynamics with grid-based docking [135].

De Novo Ligand Design

If one fails to find a drug molecule having the required interacting groups by database searching, the alternative may be to construct a ligand having active groups placed in such a way that interaction with the protein at the identified interaction sites is possible. This ligand construction process is called *de novo* ligand design. A large number of *de novo* design programs are available. These can be divided into three main categories: those that connect molecular fragments placed at the interaction sites to obtain a ligand (linking), those that start from one fragment and connect fragments sequentially to it (growing) and random connection methods. The last category includes the genetic algorithm methods. Most of the random connection methods start from an initial "pool" of fragments and construct ligands by making and breaking connections between the fragments. Molecular fragments placed at possible

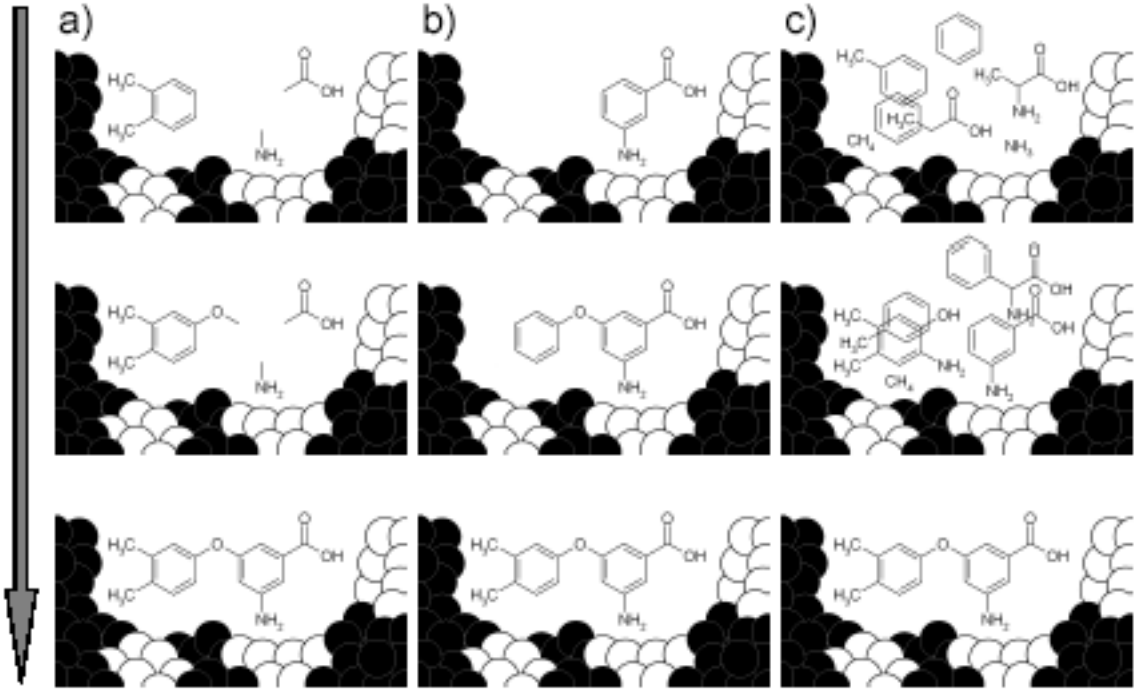


Fig. (4). Three main categories of *de novo* ligand design methods.

Black spheres indicate hydrophobic areas of the protein, while white spheres indicate hydrophilic areas. In the linking approach (a), molecular fragments placed close to important residues of the protein are connected to obtain a ligand. The growing approach (b) starts from one fragment and connects fragments sequentially to it. Most of the random connection methods (c) start from an initial "pool" of fragments and construct ligands by making and breaking connections between the fragments.

Table 2. Some De Novo Ligand Design Programs

Method	Type[b]	Ref	Url
BUILDER	L	[137]	http://thalassa.ca.sandia.gov/~dcroe/
CAVEAT	L	[138]	http://www.cchem.berkeley.edu/~pabgrp/Data/caveat.html
HOOK	L	[139]	http://www.accelrys.com/quanta/mcss_hook.html
LUDI	L	[121]	http://www.accelrys.com/insight/ludi.html
PRO_SELECT	L	[140]	http://www.protherics.com/wtech_camdt.html
SKELGEN	L	[141]	http://www.denovopharma.com/
SmoG	G	[142]	http://www-shakh.harvard.edu/~smog/
CombiSMoG	G	[143]	http://www.concurrentpharma.com/
SPLICE	L	[144]	http://www.tripos.com/
SPROUT	G	[145]	http://www.simbiosys.ca/sprout/
LigBuilder	L+G	[146]	http://mdl.ipc.pku.edu.cn/drug_design/work/ligbuilder.html
LeapFrog	G	Tripos	http://www.tripos.com/
DycoBlock	L	[147]	yyshi@iris.bio.ustc.edu.cn
ADAPT	R	[148]	http://mako.cgl.ucsf.edu/~spegg/
LEA	R	[149]	douguetl@caramail.com

[b] L – linking approach, G – growing approach, R – random connection approach

interaction sites in the receptor binding pocket found by methods such as PASSA can be used as starting points for all three approaches. These approaches are illustrated in Fig. (**4**), and Table 2 lists some *de novo* ligand design programs and the approaches they use. A more complete listing of available *de novo* ligand design methods can be found in Schneider *et al.* [136].

There are a number of limitations to existing *de novo* ligand design methods. Most of these methods do not take factors such as synthetic accessibility, bioavailability and metabolic properties into account. Many of the ligand suggestions have large and complex structures. Recently, some programs have been developed that attempt to take such factors into account. An example is LigBuilder [146], which uses a filter to make sure that the structures produced have reasonable ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties. As for molecular docking, most *de novo* ligand design methods use rigid protein structure models. Recently, some methods have been developed that attempt to take protein flexibility into account. A new version of DycoBlock, F-DycoBlock [150] uses multiple-copy stochastic molecular dynamics to account for fluctuations in the protein structure. Carlson *et al.* developed the "Dynamic Pharmacophore Method" [151], that determines pharmacophore models for a large number of MD snapshots. Protein flexibility in drug design has been reviewed by Carlson and McCammon [152,153] and Wong and McCammon [154].

Most *de novo* ligand design methods use simplified scoring functions for the ligand-receptor system to estimate binding affinity, mainly in order to speed up the calculations. Solvation effects are typically omitted. Energy-based scoring functions use molecular mechanics force fields to estimate the binding energy, while rule-based scoring functions use rules derived from analysis of structural databases. Energy-based scoring functions are slow, and sensitive to errors in the protein structure, atomic charges and protonation states. In the same way as for GRID, the force field scoring methods are often sensitive to small errors in the atomic positions. Rule-based scoring functions are often very simple, and are highly dependent on the amount of structural data used to derive the rules. In spite of these limitations, *de novo* ligand design methods have contributed to the development of several important drug leads [155], and have proved very useful when combined with some expert knowledge in medicinal chemistry. In recent years several cases of successful application of *de novo* ligand design methods have been reported, as described in the introduction. An important example is the discovery of STI-571, which is a selective inhibitor of Abl kinase, and is currently being used as a therapeutic agent against chronic myelogenous leukaemia [6,156]. Other examples include the development of antifungal agent [157] using LUDI and the design of aspartyl protease inhibitors using a growth type algorithm. The aspartyl protease inhibitors were verified experimentally [158].

CONCLUSIONS

Homology modelling has significant potential as a tool in rational drug design, in particular in high throughput *in silico* screening or simulation approaches. However, although the methods already are very useful, as demonstrated in several drug design projects, significant improvement is needed before the tools are robust and

general enough for large scale use. All aspects discussed in this review may need some improvement, but a few selected areas may benefit from some extra attention. The quality of the final structure depends mainly on the quality of the target-template alignment. Any improvement in alignment protocols will improve the final model. However, there will always be structural differences between target and templates, and these differences have to be identified and compensated for by *ab initio* modelling or by optimisation methods. In particular optimisation methods based on molecular mechanics and dynamics protocols still represent a weak point, although it is reasonable to assume that it should be possible to improve most models by using a good force field and simulation protocol. Finally, protein structures or ligands are not rigid systems, they have a high degree of flexibility, and docking or design methods that are able to take both the flexibility and small structural errors into account may give improved performance. Improvements in these and other areas may finally turn homology-based rational drug design into a really useful tool for the pharmaceutical industry.

ACKNOWLEDGEMENTS

This work has been supported by the Norwegian research council, projects 139617/140 and 138754/432.

ABBREVIATIONS

QSAR	=	Quantitative Structure-Activity Relationship.
CoMSIA	=	Comparative Molecular Similarity Indices Analysis. 3D QSAR method using Gaussian property distributions.
CoMFA	=	Comparative Molecular Field Analysis. 3D QSAR method using calculations of interaction energies between the ligands and probe atoms placed on a regular grid.
PCA	=	Principal Component Analysis. Statistical data analysis method.
DPLSR	=	Discriminant Partial Least Squares Regression. Regression method where the dependent variables are indicator variables.
GRID	=	Method for analysis of protein binding sites by calculation of interaction energies between the protein and probe atoms placed on a regular grid.
MCSS	=	Multiple Copies Simultaneous Search. Method for analysis of protein binding sites by calculation of interaction energies between the protein and probe molecules placed in the binding site.
PASSA	=	Protein Alpha Shape Similarity Analysis. Method for analysis of protein binding sites using a combination of Gaussian property distributions and DPLSR.
ADMET	=	Absorption Distribution Metabolism Excretion Toxicity

REFERENCES

[1] Greer, J.; Erickson, J.W.; Baldwin, J.J.; Varney, M.D. *J. Med. Chem.,* **1994,** *37*, 1035.
[2] Cohen, J. *Science,* **1996,** *272*, 1882.
[3] Wlodawer, A.; Vondrasek, J. *Annu. Rev. Biophys. Biomol. Struct.,* **1998,** *27*, 249.
[4] Varghese, J.N. *Drug Dev. Res.,* **1999,** *46*, 176.
[5] Gray, N.S.; Wodicka, L.; Thunnissen, A.M.; Norman, T.C.; Kwon, S.; Espinoza, F.H.; Morgan, D.O.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S.H.; Lockhart, D.J.; Schultz, P.G. *Science,* **1998,** *281*, 533.
[6] Capdeville, R.; Buchdunger, E.; Zimmermann, J.; Matter, A. *Nat. Rev. Drug Discov.,* **2002,** *1*, 493.
[7] Davies, T.G.; Tunnah, P.; Meijer, L.; Marko, D.; Eisenbrand, G.; Endicott, J.A.; Noble, M.E. *Structure (Camb.),* **2001,** *9*, 389.
[8] Ghosh, S.; Liu, X.P.; Zheng, Y.; Uckun, F.M. *Curr. Cancer Drug Targets,* **2001,** *1*, 129.
[9] Zhu, X.; Kim, J.L.; Newcomb, J.R.; Rose, P.E.; Stover, D.R.; Toledo, L.M.; Zhao, H.; Morgenstern, K.A. *Structure Fold. Des.,* **1999,** *7*, 651.
[10] Sawyer, T.; Boyce, B.; Dalgarno, D.; Iuliucci, J. *Expert. Opin. Investig. Drugs,* **2001,** *10*, 1327.
[11] Govindarajan, S.; Recabarren, R.; Goldstein, R.A. *Proteins,* **1999,** *35*, 408.
[12] Enyedy, I.J.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X.; Cao, Y.; Guo, R.; Li, B.; Zhu, X.; Huang, Y.; Long, Y.Q.; Roller, P.P.; Yang, D.; Wang, S. *J. Med. Chem.,* **2001,** *44*, 4313.
[13] Furet, P.; Zimmermann, J.; Capraro, H.G.; Meyer, T.; Imbach, P. *J. Comput. Aided Mol. Des.,* **2000,** *14*, 403.
[14] Schapira, M.; Raaka, B.M.; Samuels, H.H.; Abagyan, R. *Proc. Natl. Acad. Sci. U. S. A.,* **2000,** *97*, 1008.
[15] Sabnis, Y.; Rosenthal, P.J.; Desai, P.; Avery, M.A. *J. Biomol. Struct. Dyn.,* **2002,** *19*, 765.
[16] Schafferhans, A.; Klebe, G. *J. Mol. Biol.,* **2001,** *307*, 407.
[17] Tøndel, K.; Anderssen, E.; Drabløs, F. *J. Comput. Aided Mol. Des.,* **2002,** *16*, 831.
[18] Lesk, A.M.; Chothia, C. *J. Mol. Biol.,* **1980,** *136*, 225.
[19] Chothia, C.; Lesk, A.M. *EMBO J.,* **1986,** *5*, 823.
[20] Raha, K.; Wollacott, A.M.; Italia, M.J.; Desjarlais, J.R. *Protein Sci.,* **2000,** *9*, 1106.
[21] Zvelebil, M.J.; Barton, G.J.; Taylor, W.R.; Sternberg, M.J. *J. Mol. Biol.,* **1987,** *195*, 957.
[22] Schonbrun, J.; Wedemeyer, W.J.; Baker, D. *Curr. Opin. Struct. Biol.,* **2002,** *12*, 348.
[23] Al Lazikani, B.; Jung, J.; Xiang, Z.; Honig, B. *Curr. Opin. Chem. Biol.,* **2001,** *5*, 51.
[24] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. *J. Mol. Biol.,* **1990,** *215*, 403.
[25] Pearson, W.R.; Lipman, D.J. *Proc. Natl. Acad. Sci. U. S. A.,* **1988,** *85*, 2444.
[26] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. *Nucleic Acids Res.,* **1997,** *25*, 3389.
[27] Edwards, Y.J.; Cottage, A. *Mol. Biotechnol.,* **2003,** *23*, 139.
[28] Lundstrom, J.; Rychlewski, L.; Bujnicki, J.; Elofsson, A. *Protein Sci.,* **2001,** *10*, 2354.
[29] Venclovas, C. *Proteins,* **2001,** *Suppl 5*, 47.
[30] Qian, B.; Goldstein, R.A. *Proteins,* **2002,** *48*, 605.
[31] Needleman, S.B.; Wunsch, C.D. *J. Mol. Biol.,* **1970,** *48*, 443.
[32] Smith, T.F.; Waterman, M.S. *J. Mol. Biol.,* **1981,** *147*, 195.
[33] Jaroszewski, L.; Rychlewski, L.; Godzik, A. *Protein Sci.,* **2000,** *9*, 1487.
[34] Thompson, J.D.; Gibson, T.J.; Plewniak, F.; Jeanmougin, F.; Higgins, D.G. *Nucleic Acids Res.,* **1997,** *25*, 4876.
[35] Lee, C.; Grasso, C.; Sharlow, M.F. *Bioinformatics,* **2002,** *18*, 452.
[36] Morgenstern, B.; Frech, K.; Dress, A.; Werner, T. *Bioinformatics,* **1998,** *14*, 290.
[37] Morgenstern, B. *Bioinformatics,* **1999,** *15*, 211.
[38] Notredame, C.; Higgins, D.G.; Heringa, J. *J. Mol. Biol.,* **2000,** *302*, 205.
[39] Lassmann, T.; Sonnhammer, E.L. *FEBS Lett.,* **2002,** *529*, 126.
[40] Elofsson, A. *Proteins,* **2002,** *46*, 330.
[41] Lambert, C.; Leonard, N.; De, B., X; Depiereux, E. *Bioinformatics,* **2002,** *18*, 1250.
[42] Karwath, A.; King, R.D. *BMC Bioinformatics,* **2002,** *3*, 11.
[43] Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. *Nucleic Acids Res.,* **2002,** *30*, 3059.

[44] Blake, J.D.; Cohen, F.E. *J. Mol. Biol.,* **2001,** *307*, 721.
[45] Cristobal, S.; Zemla, A.; Fischer, D.; Rychlewski, L.; Elofsson, A. *BMC Bioinformatics,* **2001,** *2*, 5.
[46] Yang, A.S. *Bioinformatics,* **2002,** *18*, 1658.
[47] Holm, L.; Sander, C. *J. Mol. Biol.,* **1993,** *233*, 123.
[48] Gerstein, M.; Levitt, M. *Protein Sci.,* **1998,** *7*, 445.
[49] Singh, A.P.; Brutlag, D.L. *Proc. Int. Conf. Intell. Syst. Mol. Biol.,* **1997,** *5*, 284.
[50] Deane, C.M.; Kaas, Q.; Blundell, T.L. *Bioinformatics,* **2001,** *17*, 541.
[51] Koehl, P.; Levitt, M. *Nat. Struct. Biol.,* **1999,** *6*, 108.
[52] Guex, N.; Peitsch, M.C. *Electrophoresis,* **1997,** *18*, 2714.
[53] Vriend, G. *J. Mol. Graph.,* **1990,** *8*, 52.
[54] Blundell, T.; Carney, D.; Gardner, S.; Hayes, F.; Howlin, B.; Hubbard, T.; Overington, J.; Singh, D.A.; Sibanda, B.L.; Sutcliffe, M. *Eur. J. Biochem.,* **1988,** *172*, 513.
[55] Johnson, M.S.; Srinivasan, N.; Sowdhamini, R.; Blundell, T.L. *Crit Rev. Biochem. Mol. Biol.,* **1994,** *29*, 1.
[56] Sali, A.; Blundell, T.L. *J. Mol. Biol.,* **1993,** *234*, 779.
[57] Sutcliffe, M.J.; Haneef, I.; Carney, D.; Blundell, T.L. *Protein Eng.,* **1987,** *1*, 377.
[58] Sutcliffe, M.J.; Hayes, F.R.; Blundell, T.L. *Protein Eng.,* **1987,** *1*, 385.
[59] Bruccoleri, R.E. *Mol. Sim.,* **1993,** *10*, 151.
[60] Greer, J. *Proteins,* **1990,** *7*, 317.
[61] Abagyan, R.; Batalov, S.; Cardozo, T.; Totrov, M.; Webber, J.; Zhou, Y. *Proteins,* **1997,** *Suppl 1*, 29.
[62] Levitt, M. *J. Mol. Biol.,* **1992,** *226*, 507.
[63] Guex, N.; Diemand, A.; Peitsch, M.C. *Trends Biochem. Sci.,* **1999,** *24*, 364.
[64] Peitsch, M.C. *Biochem. Soc. Trans.,* **1996,** *24*, 274.
[65] Kolodny, R.; Koehl, P.; Guibas, L.; Levitt, M. *J. Mol. Biol.,* **2002,** *323*, 297.
[66] Sali, A.; Potterton, L.; Yuan, F.; van Vlijmen, H.; Karplus, M. *Proteins,* **1995,** *23*, 318.
[67] de Bakker, P.I.; DePristo, M.A.; Burke, D.F.; Blundell, T.L. *Proteins,* **2003,** *51*, 21.
[68] Fiser, A.; Do, R.K.; Sali, A. *Protein Sci.,* **2000,** *9*, 1753.
[69] Galaktionov, S.; Nikiforovich, G.V.; Marshall, G.R. *Biopolymers,* **2001,** *60*, 153.
[70] Chinea, G.; Padron, G.; Hooft, R.W.; Sander, C.; Vriend, G. *Proteins,* **1995,** *23*, 415.
[71] Dunbrack, R.L., Jr.; Cohen, F.E. *Protein Sci.,* **1997,** *6*, 1661.
[72] Dunbrack, R.L., Jr. *Curr. Opin. Struct. Biol.,* **2002,** *12*, 431.
[73] Liu, Z.; Jiang, L.; Gao, Y.; Liang, S.; Chen, H.; Han, Y.; Lai, L. *Proteins,* **2003,** *50*, 49.
[74] Yang, J.M.; Tsai, C.H.; Hwang, M.J.; Tsai, H.K.; Hwang, J.K.; Kao, C.Y. *Protein Sci.,* **2002,** *11*, 1897.
[75] Martin, A.C.; MacArthur, M.W.; Thornton, J.M. *Proteins,* **1997,** *Suppl 1*, 14.
[76] Marti-Renom, M.A.; Stuart, A.C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.,* **2000,** *29*, 291.
[77] Mezei, M. *Protein Eng.,* **1998,** *11*, 411.
[78] Lessel, U.; Schomburg, D. *Proteins,* **1999,** *37*, 56.
[79] Wohlfahrt, G.; Hangoc, V.; Schomburg, D. *Proteins,* **2002,** *47*, 370.
[80] Charifson, P.S. *Practical application of computer-aided drug design*, Marcel Dekker, Inc.: New York, **1997**.
[81] Flohil, J.A.; Vriend, G.; Berendsen, H.J. *Proteins,* **2002,** *48*, 593.
[82] Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. *Proteins,* **2001,** *Suppl 5*, 2.
[83] Eyrich, V.A.; Marti-Renom, M.A.; Przybylski, D.; Madhusudhan, M.S.; Fiser, A.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B. *Bioinformatics,* **2001,** *17*, 1242.
[84] Marti-Renom, M.A.; Madhusudhan, M.S.; Fiser, A.; Rost, B.; Sali, A. *Structure (Camb.),* **2002,** *10*, 435.
[85] Sanchez, R.; Sali, A. *Proc. Natl. Acad. Sci. U. S. A.,* **1998,** *95*, 13597.
[86] Dodson, E.J.; Davies, G.J.; Lamzin, V.S.; Murshudov, G.N.; Wilson, K.S. *Structure Fold. Des.,* **1998,** *6*, 685.
[87] MacArthur, M.W.; Laskowski, R.A.; Thornton, J.M. *Curr. Opin. Struct. Biol.,* **1994,** *4*, 731.
[88] Ramachandran, G.N.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.,* **1963,** *7*, 95.
[89] Laskowski, R.A.; MacArthur, M.W.; Moss, D.S.; Thornton, J.M. *J. Appl. Cryst.,* **1993,** *26*, 283.
[90] Bujnicki, J.; Rychlewski, L.; Fischer, D. *Bioinformatics,* **2002,** *18*, 1391.
[91] Sippl, M.J. *Proteins,* **1993,** *17*, 355.
[92] Luthy, R.; Bowie, J.U.; Eisenberg, D. *Nature,* **1992,** *356*, 83.
[93] Godzik, A.; Kolinski, A.; Skolnick, J. *Protein Sci.,* **1995,** *4*, 2107.
[94] Godzik, A. *Structure,* **1996,** *4*, 363.
[95] Jaroszewski, L.; Pawlowski, K.; Godzik, A. *J. Mol. Model.,* **1998,** *4*, 294.
[96] Pawlowski, K.; Jaroszewski, L.; Bierzynski, A.; Godzik, A. *Pac. Symp. Biocomput.,* **1997,** 328.
[97] Siew, N.; Elofsson, A.; Rychiewski, L.; Fischer, D. *Bioinformatics,* **2000,** *16*, 776.
[98] Smith, A. *Nat. Rev. Drug Discov.,* **2002,** *1*, 3.
[99] Sotriffer, C.; Klebe, G. *Farmaco,* **2002,** *57*, 243.
[100] Miranker, A.; Karplus, M. *Proteins,* **1991,** *11*, 29.
[101] Stultz, C.M.; Karplus, M. *Proteins,* **1999,** *37*, 512.
[102] Goodford, P.J. *J. Med. Chem.,* **1985,** *28*, 849.
[103] Wade, R.C.; Clark, K.J.; Goodford, P.J. *J. Med. Chem.,* **1993,** *36*, 140.
[104] Wade, R.C.; Goodford, P.J. *J. Med. Chem.,* **1993,** *36*, 148.
[105] Powers, R.A.; Shoichet, B.K. *J. Med. Chem.,* **2002,** *45*, 3222.
[106] Powers, R.A.; Morandi, F.; Shoichet, B.K. *Structure (Camb.),* **2002,** *10*, 1013.
[107] von Itzstein, M.; Wu, W.Y.; Kok, G.B.; Pegg, M.S.; Dyason, J.C.; Jin, B.; Van Phan, T.; Smythe, M.L.; White, H.F.; Oliver, S.W.; . *Nature,* **1993,** *363*, 418.
[108] von Itzstein, M.; Dyason, J.C.; Oliver, S.W.; White, H.F.; Wu, W.Y.; Kok, G.B.; Pegg, M.S. *J. Med. Chem.,* **1996,** *39*, 388.
[109] English, A.C.; Groom, C.R.; Hubbard, R.E. *Protein Eng.,* **2001,** *14*, 47.
[110] Kastenholz, M.A.; Pastor, M.; Cruciani, G.; Haaksma, E.E.; Fox, T. *J. Med. Chem.,* **2000,** *43*, 3033.
[111] Matter, H.; Schwab, W. *J. Med. Chem.,* **1999,** *42*, 4506.
[112] Ridderstrom, M.; Zamora, I.; Fjellstrom, O.; Andersson, T.B. *J. Med. Chem.,* **2001,** *44*, 4072.
[113] Liang, J.; Edelsbrunner, H.; Woodward, C. *Protein Sci.,* **1998,** *7*, 1884.
[114] Bohm, M.; Stürzebecher, J.; Klebe, G. *J. Med. Chem.,* **1999,** *42*, 458.
[115] Miller, M.A. *Nat. Rev. Drug Discov.,* **2002,** *1*, 220.
[116] Allen, F.H.; Kennard, O.; Taylor, R. *Acc. Chem. Res.,* **1983,** *16*, 146.
[117] Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. *J. Comput. Chem.,* **1998,** *19*, 1639.
[118] Ewing, T.J.A.; Kuntz, I.D. *J. Comput. Chem.,* **1997,** *18*, 1175.
[119] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.,* **1996,** *261*, 470.
[120] Jones, G.; Willett, P.; Glen, R.C. *J. Mol. Biol.,* **1995,** *245*, 43.
[121] Bohm, H.J. *J. Comput. Aided Mol. Des.,* **1992,** *6*, 61.
[122] Rarey, M.; Lengauer, T. *Persp. Drug Disc. Des.,* **2000,** *20*, 63.
[123] McGann, M.R.; Almond, H.R.; Nicholls, A.; Grant, J.A.; Brown, F.K. *Biopolymers,* **2003,** *68*, 76.
[124] Wojciechowski, M.; Skolnick, J. *J. Comput. Chem.,* **2002,** *23*, 189.
[125] Taylor, R.D.; Jewsbury, P.J.; Essex, J.W. *J. Comput. Aided Mol. Des.,* **2002,** *16*, 151.
[126] Lyne, P.D. *Drug Discov. Today,* **2002,** *7*, 1047.
[127] Bajorath, J. *Nat. Rev. Drug Discov.,* **2002,** *1*, 882.
[128] Leach, A.R. *J. Mol. Biol.,* **1994,** *235*, 345.
[129] Kairys, V.; Gilson, M.K. *J. Comput. Chem.,* **2002,** *23*, 1656.
[130] Anderson, A.C.; O'Neil, R.H.; Surti, T.S.; Stroud, R.M. *Chem. Biol.,* **2001,** *8*, 445.
[131] Osterberg, F.; Morris, G.M.; Sanner, M.F.; Olson, A.J.; Goodsell, D.S. *Proteins,* **2002,** *46*, 34.
[132] Lin, J.H.; Perryman, A.L.; Schames, J.R.; McCammon, J.A. *J. Am. Chem. Soc.,* **2002,** *124*, 5632.
[133] Lin, J.H.; Perryman, A.L.; Schames, J.R.; McCammon, J.A. *Biopolymers,* **2003,** *68*, 47.
[134] Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. *J. Mol. Biol.,* **2001,** *308*, 377.
[135] Broughton, H.B. *J. Mol. Graph. Model.,* **2000,** *18*, 247.
[136] Schneider, G.; Bohm, H.J. *Drug Discov. Today,* **2002,** *7*, 64.
[137] Roe, D.C.; Kuntz, I.D. *J. Comput. Aided Mol. Des.,* **1995,** *9*, 269.
[138] Lauri, G.; Bartlett, P.A. *J. Comput. Aided Mol. Des.,* **1994,** *8*, 51.
[139] Eisen, M.B.; Wiley, D.C.; Karplus, M.; Hubbard, R.E. *Proteins,* **1994,** *19*, 199.

[140] Murray, C.W.; Clark, D.E.; Auton, T.R.; Firth, M.A.; Li, J.; Sykes, R.A.; Waszkowycz, B.; Westhead, D.R.; Young, S.C. *J. Comput. Aided Mol. Des.,* **1997,** *11*, 193.
[141] Todorov, N.P.; Dean, P.M. *J. Comput. Aided Mol. Des.,* **1997,** *11*, 175.
[142] DeWitte, R.S.; Ishchenko, A.V.; Shakhnovich, E.I. *J. Am. Chem. Soc.,* **1997,** *119*, 4608.
[143] Grzybowski, B.A.; Ishchenko, A.V.; Shimada, J.; Shakhnovich, E.I. *Acc. Chem. Res.,* **2002,** *35*, 261.
[144] Ho, C.M.; Marshall, G.R. *J. Comput. Aided Mol. Des.,* **1993,** *7*, 3.
[145] Gillet, V.; Johnson, A.P.; Mata, P.; Sike, S.; Williams, P. *J. Comput. Aided Mol. Des.,* **1993,** *7*, 127.
[146] Wang, R.X.; Gao, Y.; Lai, L.H. *J. Mol. Model.,* **2000,** *6*, 498.
[147] Liu, H.; Duan, Z.; Luo, Q.; Shi, Y. *Proteins,* **1999,** *36*, 462.
[148] Pegg, S.C.; Haresco, J.J.; Kuntz, I.D. *J. Comput. Aided Mol. Des.,* **2001,** *15*, 911.
[149] Douguet, D.; Thoreau, E.; Grassy, G. *J. Comput. Aided Mol. Des.,* **2000,** *14*, 449.
[150] Zhu, J.; Fan, H.; Liu, H.; Shi, Y. *J. Comput. Aided Mol. Des.,* **2001,** *15*, 979.
[151] Carlson, H.A.; Masukawa, K.M.; McCammon, J.A. *J. Phys. Chem. A,* **1999,** *103*, 10213.
[152] Carlson, H.A.; McCammon, J.A. *Mol. Pharmacol.,* **2000,** *57*, 213.
[153] Carlson, H.A. *Curr. Opin. Chem. Biol.,* **2002,** *6*, 447.
[154] Wong, C.F.; McCammon, J.A. *Annu. Rev. Pharmacol. Toxicol.,* **2003,** *43*, 31.
[155] Sawyer, T.K. *Biotechniques,* **2001,** *31*, 1164, 1166, 1168.
[156] Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, W.T.; Clarkson, B.; Kuriyan, J. *Science,* **2000,** *289*, 1938.
[157] Ji, H.; Zhang, W.; Zhang, M.; Kudo, M.; Aoyama, Y.; Yoshida, Y.; Sheng, C.; Song, Y.; Yang, S.; Zhou, Y.; Lu, J.; Zhu, J. *J. Med. Chem.,* **2003,** *46*, 474.
[158] Ripka, A.S.; Satyshur, K.A.; Bohacek, R.S.; Rich, D.H. *Org. Lett.,* **2001,** *3*, 2309.

Paper IV

Journal of Theoretical and Computational Chemistry, Vol. 2, No. 1 (2003) 43–56

World Scientific
www.worldscientific.com

COMPUTATIONAL ANALYSIS OF THE INTERACTIONS BETWEEN THE ANGIOGENESIS INHIBITOR PD173074 AND FIBROBLAST GROWTH FACTOR RECEPTOR 1

KRISTIN TØNDEL
Department of Chemistry, Physical Chemistry,
Norwegian University of Science and Technology,
Sem Selands v. 14, N-7591 Trondheim, Norway

CHUNG F. WONG*, and J. A. MCCAMMON†
Howard Hughes Medical Institute,
Department of Pharmacology,
†*Department of Chemistry and Biochemistry,*
University of California, San Diego, 9500 Gilman Drive,
La Jolla, CA 92093-0365, USA
**c4wong@ucsd.edu*

Received 7 November 2002
Accepted 13 December 2002

We have carried out computational sensitivity analysis to analyze the interactions between the inhibitor PD173074 and FGFR1 in order to identify the determinants of their recognition and generate insights into further refining the inhibitor. The analysis has identified the parts of the inhibitor that are already useful for binding, e.g. the part that recognizes the linker connecting the N-terminal and C-terminal lobes of the kinase domain. These parts are profitably kept during a lead optimization process. The analysis has also pointed out regions of the inhibitors that may be useful to modify to improve its binding affinity, e.g. the dimethoxyphenyl ring. Comparative structural analysis of the binding pocket of almost 400 protein kinases also suggests that modifying the dimethoxyphenyl moiety might improve selective binding. Selectivity may be achieved not only by introducing groups to the 3 and 5 positions but also to the 1 and 6 positions. Replacing the tertiary amines by hydrocarbon might also improve binding affinity.

Keywords: Computational sensitivity analysis; continuum-solvent binding energy calculations; comparative sequence/structure analysis.

1. Introduction

Angiogenesis (the biological process by which new capillaries are formed from preexisting vessels) is involved in embryo development, ovulation, and wound repair. It is also essential for growth and metastasis of tumors.[1,2] Pathological angiogenesis (abnormal rapid proliferation of blood vessels) is involved in a large number of other diseases as well, such as diabetic retinopathy, atherosclerosis, rheumatoid arthritis, age-related macular degeneration and psoriasis.[3–8] Hence, angiogenic factors and their receptors are common targets for development of therapeutic agents.

The normal regulation of angiogenesis is governed by a fine balance between factors that induce the formation of blood vessels and those that halt or inhibit the process.[9] Numerous factors that regulate

angiogenesis have been identified, including members of the fibroblast growth factor (FGF) family, vascular endothelial growth factor (VEGF), angiogenin, transforming growth factor (TGF) α and β, platelet-derived growth factor (PDGF), platelet-derived endothelial cell growth factor (PDECGF), tumor necrosis factor (TNF) α, interleukins, chemokines and angiopoietins.[9,10]

The effects of the potent angiogenic factors FGF and VEGF are mediated through cell surface receptors (fibroblast growth factor receptor (FGFR) and vascular endothelial growth factor receptor (VEGFR)) that possess intrinsic protein tyrosine kinase activity.[11] FGFR and VEGFR are members of the receptor tyrosine kinase (RTK) family of enzymes. RTKs consist of an extracellular portion that binds polypeptide ligands, a transmembrane helix, and a cytoplasmic portion that possesses tyrosine kinase catalytic activity.[12] The majority of RTKs are monomeric in the absence of ligand. Binding of ligand to RTKs leads to receptor oligomerization and tyrosine autophosphorylation. Autophosphorylation of tyrosine residues leads to increased kinase catalytic activity, and generation of docking sites for protein substrates. The RTKs catalyze the transfer of the γ phosphate of ATP to the hydroxyl group of a tyrosine in a substrate protein. This triggers signaling cascades that participate in a large number of biological processes.

Mohammadi *et al.*[11] have reported the crystal structure of a compound of the pyrido[2,3-d]pyrimidine class (PD173074) that selectively inhibits the tyrosine kinase activity of FGFR and VEGFR in complex with the FGF receptor tyrosine kinase domain. This inhibitor contains a dimethoxy phenyl group that occupies a pocket in the ATP-binding cleft that is not occupied by ATP. Mohammadi *et al.*[11] suggest that this group is important for the selective binding of this inhibitor. In this paper we have investigated this hypothesis further by carrying out computational sensitivity analysis and by comparative analysis of the binding pocket of almost 400 protein kinases.

The basic idea of computational sensitivity analysis is similar to genetic experiments, in which one examines whether a particular feature of an amino acid is affecting the property of a protein by mutating the amino acid into another one that no longer contains the feature. In a computational sensitivity analysis, one "mutates" parameters of a molecular model, such as atomic partial charges and dipole moments of functional groups to examine the significance of these features in affecting binding affinity. This analysis can be done efficiently by using mathematical methods that are widely used by engineers.[13] The study here further examines the kind of information sensitivity analysis can provide in aiding the design of therapeutic agents. In this work, we use relatively inexpensive implicit-solvent models instead of the explicit-solvent molecular dynamics simulation models employed earlier.[13] We use this approach to identify the parts of a drug lead that are most significant for binding and the parts that should be modified to improve binding affinity. With this knowledge, one can generate useful insights into the optimization of a drug lead. For example, one may want to keep the parts that are already useful and focus on modifying those parts that are not yet beneficial. This information can also add useful constraints to designing focused chemical libraries that may produce more useful new hits. The parts of a drug lead that have been identified to be useful for binding can also guide the construction of pharmacophore models for mining new drug leads from small-molecule libraries. We previously applied such an approach to studying the binding of balanol and a peptide inhibitor to protein kinase A.[14,15] Here we extend this study to elucidating the interactions between PD173074 and FGFR1. Again, we focus on studying electrostatics determinants here. Studying other determinants such as the size of functional groups may require proper account of ligand and receptor flexibility. Since this work focuses on examining the feasibility of using fixed-conformation models for speedier initial evaluations, we avoid examining determinants that the fixed conformation models may not be sufficiently reliable to address. More general molecular dynamics simulation model can take flexibility into account but with a substantial increase in computational costs.[13]

In designing drugs targeting protein kinases, specificity is also an important factor to consider because many promising drug leads bind to the ATP-binding pocket, which is a common feature of all protein kinases. We previously showed that one could obtain useful insights into specificity by combining structural information obtained from protein crystallography and sequence information of several hundred

protein kinases.[14,15] This idea is further applied here to generate insights into how PD173074 may be modified to improve selectivity.

2. Methods

2.1. *Computational sensitivity analysis*

The crystal structure of the FGF receptor tyrosine kinase domain in complex with the angiogenesis inhibitor PD173074 was obtained from the Protein Data Bank (PDB) (entry 2FGI).[16] Hydrogen atoms for the protein and the ligand were added by using the CHARMM[17] and QUANTA[18] packages respectively. The atomic charges of the inhibitor were obtained by using the Merz-Kollman method[19,20] in Gaussian98.[21] The 6-31G* basis set was used. Since the united-atom representation was used for the nonpolar hydrogens in the binding energy calculations, the atomic charges for the hydrogens were added to the atomic charges of the heavy atoms to which they attach. The hydrogens of the proteins were first relaxed by 200 steps of steepest descent energy minimization keeping the heavy atoms of the protein and all the ligand atoms fixed. The ligand atoms were then relaxed by 200 steps of steepest descent energy minimization with the whole protein held fixed. The CHARMm22 forcefield[22] were used for the energy minimization.

The UHBD program[23,24] was used for the binding energy calculations. The free energy, G, for the complex, the protein, or the ligand, was calculated according to Eq. (1).

$$G = G_{\text{solv}} + G_{\text{coul}} + G_{\text{surf}} \tag{1}$$

in which G_{solv} is the electrostatics contribution to the solvation energy obtained by solving the Poisson equation and G_{coul} is the Coulombic energy calculated by Coulomb's law using the dielectric constant of the solute. G_{surf} describes hydrophobic contributions estimated by multiplying the solvent accessible surface area (determined by using a probe sphere of 1.4 Å radius) by 0.025 kcal/mol/Å^2. In these calculations, the atomic charges and radii from CHARMm22[22] were used except that the atomic charges of the ligand were obtained by the quantum calculations described above. In solving the Poisson equation, the dielectric constant of the solvent was 78 and the dielectric constant for the solute interior was 2. The size of the grid was 240 Å × 240 Å × 240 Å and the grid spacing was 0.3 Å. The binding energy of the complex was estimated by

$$\Delta G_{\text{binding}} = G_{\text{bin}} - (G_{\text{apo}} + G_{\text{lig}}) \tag{2}$$

where G_{bin}, G_{apo} and G_{lig} are the free energy of the complex, the protein, and the ligand respectively.

Parallel to previous sensitivity analysis based on molecular dynamics simulation, we gauge the significance of a model parameter in affecting a binding energy by calculating derivatives of the form $dG/d\lambda_i$. In this work, we estimated these derivatives by

$$d\Delta G/d\lambda_i = (\Delta G_{\text{binding}}(\text{mutant}) - \Delta G_{\text{binding}}(\text{wildtype}))/\Delta\lambda_i \tag{3}$$

where $\Delta G_{\text{binding}}$(wildtype) is the binding free energy between PD173074 and FGFR1, and $\Delta G_{\text{binding}}$(mutant) is the corresponding quantity when the collective charge or dipole moment of an atom or a functional group is changed by $\Delta\lambda_i$. The collective charge equals the charge of an atom if only one atom is involved or the sum of the charges within a group if a group of atoms is involved. We also scaled a derivative by $\Delta\lambda_i$ to gauge the effect of turning on the collective charge or dipole moment of an atom or a group of atoms from zero. In other words, we calculated $d\Delta G/d\lambda_i\lambda_i$ where λ_i represents the collective charge or dipole moment of an atom or a group of atoms. A negative $d\Delta G/d\lambda_i\lambda$ signifies an improvement of the binding affinity when the collective charge or dipole moment is turned on.

We also estimated the significance of pairwise interactions on affecting binding by calculating second order derivatives of the form $d^2\Delta G/d\lambda_i d\lambda_j)\lambda_i\lambda_j$. These second order derivatives measure pairwise interactions as in double mutagenesis experiments. The significance of a pairwise interaction is estimated by subtracting the change in binding free energy resulting from two single point mutations from the change resulting from making the two mutations simultaneously. As in a previous work,[15] we calculated the second derivatives or pairwise interactions in the same way as the experimentalists except that charges or dipole moments, instead of amino acids, were changed.

The inhibitor studied in this project contains a nitrogen (N1) that might be protonated in solution near neutral pH. However, this nitrogen may not be protonated inside the protein if its environment is not

sufficiently polar. To check whether the protonated form is preferred in this protein, we used the following thermodynamic cycle to calculate the free energy change ΔG that gives us this information:

$$\begin{array}{ccc} \mathrm{LH}^+ & \xrightarrow{\Delta G_2} & \mathrm{L} + \mathrm{H}^+ \\ \Delta G_1 \downarrow & & \downarrow \Delta G_3 \\ \mathrm{P:LH}^+ & \xrightarrow[\Delta G]{} & \mathrm{PL} + \mathrm{H}^+ \end{array}$$

LH: Protonated ligand
L: Unprotonated ligand
P:LH: Protonated ligand bound to protein
P:L: Unprotonated ligand bound to protein

such that

$$\Delta G = \Delta G_3 + \Delta G_2 - \Delta G_1 \quad (4)$$

ΔG_2 was estimated by using the measured pK_a of triethylamine in water at 25°C ($\Delta G_2 = -RT \ln K_a$).[25] ΔG_1 was estimated by calculating the difference between the free energy of the complex in which the ligand was protonated, G_{bin}(protonated ligand), and that of the protonated form of the ligand in solution, G_{ligand}(protonated):

$$\Delta G_1 = G_{\mathrm{bin}}(\text{protonated ligand}) - G_{\mathrm{ligand}}(\text{protonated}) \quad (5)$$

G_{bin}(protonated ligand) and G_{ligand}(protonated) were calculated according to Eq. (1). The protonated ligand in solution was only modeled as protonated triethylamine because we used the experimental pK_a of triethylamine to complete the thermodynamic cycle. The coordinates for triethylamine were obtained by deleting all the other atoms of the ligand PD173074. ΔG_3 was estimated by

$$\Delta G_3 = G_{\mathrm{bin}}(\text{neutral ligand}) - G_{\mathrm{ligand}}(\text{neutral}) \quad (6)$$

where G_{bin}(neutral ligand) and G_{ligand}(neutral) were respectively the free energy of the complex in which the ligand was neutral and the free energy of the neutral ligand in solution. These terms were again calculated according to Eq. (1). The neutral ligand was modeled by setting the charge of the ligand atoms corresponding to those of triethylamine (i.e. C1, C2, C3, C4, C5, C6, N1 and H4) to zero. These changes reduced the net charge of the ligand from +1 to zero.

2.2. *Desolvation effect*

Desolvation of a charge/polar group can have a significant impact on binding free energy. In this work, we estimated the desolvation effect of such a group by calculating the binding free energy with all charges except those associated with the group of interest to zero. This way, the contributions from the interactions of the group with other atoms in the system were eliminated and the calculated free energy change only reflects the desolvation of the group.

2.3. *Prediction of binding affinities*

We first tested the ability of our computational model to predict binding affinity by comparing the calculated results with experimental binding affinities to FGFR1. Eight different ligands, including the angiogenesis inhibitor PD173074, were included in the study. Once the model was validated, we used the model for sensitivity analysis. The insights provided by sensitivity analysis were tested by making suitable derivatives of PD173074 and calculating their binding affinity to FGFR1. In cases where new functional groups were introduced, the structure of the ligand in complex with FGFR1 was energy optimized to a root-mean-square gradient of 0.01 kcal/mol/Å^2 using the molecular mechanics force field MMFF94[26] together with the PB/SA model implemented in the MOE package.[27] A smooth non-bonded cutoff was used in the range 10–12 Å. All atoms of the complex except those of the introduced functional group were kept fixed during energy minimization.

2.4. *Analysis of variability of amino acid types at various sites of the protein-ligand interface*

To gain insights into which parts of the "lead compound" PD173074 one should focus on modifying to achieve selectivity, we have analyzed the distribution of amino acid types at various sites near the protein-ligand interface. A database of almost 400 protein kinases was analyzed. This analysis should be more insightful than many others

that included only a few protein kinases. Here, we used the database of protein kinases whose sequences were aligned by Hanks and Quinn[28] and is available in the Protein Kinase Resource maintained at the University of California, San Diego (http://www.sdsc.edu/kinases/pkr/pk_catalytic/pk_hanks_seq_align_long.html). The database also includes protein kinases found in species other than human. These were included in the analysis since there are protein kinases in the human genome that have not yet been identified or included in this database. We used protein kinases from other species to approximate the amino acid composition of the binding pocket of human protein kinases that are not yet in the database. The analysis can be easily repeated with a more complete database of human protein kinases when it becomes available.

3. Results and Discussions

3.1. *Prediction of binding affinities*

We first checked the computational model used in this work by applying it to calculate the binding affinity for eight ligands (including PD173074) with known IC_{50}. The structure of the ligands is shown in Fig. 1 and the results are given in Table 1. In Fig. 2, the calculated $\Delta G_{\text{binding}}$ is plotted against $\log(IC_{50})$. The correlation between simulated and experimental results is quite good with a correlation coefficient of 0.8. This gives us confidence on applying the computational model to identify the determinants of protein-ligand recognition and to generate hints for further refining PD173074.

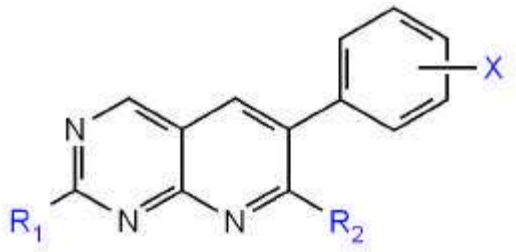


Fig. 1. Basic scaffold of the ligands studied in this work.

3.2. *Computational sensitivity analysis*

Figure 3(a) shows the angiogenesis inhibitor used in this study (PD173074) along with atom labels to facilitate later discussions. The charges used for the binding energy calculations are shown in Fig. 3(b). Some of the atomic charges obtained from the Merz–Kollman method[19] are somewhat large. Methods imposing charge constraints to reduce the magnitude of atomic charges, such as RESP,[29] can be used. However, it is not essential to do this here because the charges were largely used for describing the electrostatic field around the ligand for fixed-conformation binding energy calculations, not for conformational sampling in which intramolecular interactions among the ligand atoms also needed to be more suitably modeled.

Table 1. Calculated binding energies of eight different ligands to FGFR1 and their comparison to experimental $\log(IC_{50})$ values.

Ligand	X	R_1	R_2	IC_{50}	Log (IC_{50})	$\Delta G_{\text{binding}}$ (kcal/mol)
1	$3',5'$-$(OMe)_2$	$NH(CH_2)_4NEt_2$	NHCONHtBu	21.5[a]	1.33	−20.32
2	$2',6'$-$(Cl)_2$	NH_2	NHCONHtBu	130[b]	2.11	−15.70
3	$3,5'$-$(OMe)_2$	NH_2	NHCONHtBu	60[b]	1.78	−16.97
4	$2',6'$-$(Cl)_2$	$NH(CH_2)_4NEt_2$	NHCONHEt	49[c]	1.69	−17.55
5	$2',6'$-$(Cl)_2$	$NH(CH_2)_4NEt_2$	NHCONHtBu	48[c]	1.68	−18.32
6	H	NH_2	NHCONHtBu	3700[d]	3.57	−15.70
7	$2',6'$-$(Cl)_2$	NH_2	NH_2	3000[d]	3.48	−12.09
8	$3',5'$-$(OMe)_2$	NH_2	NH_2	230[d]	2.36	−14.08

[a]Ref. 11.
[b]Ref. 33.
[c]Ref. 34.
[d]Ref. 35.

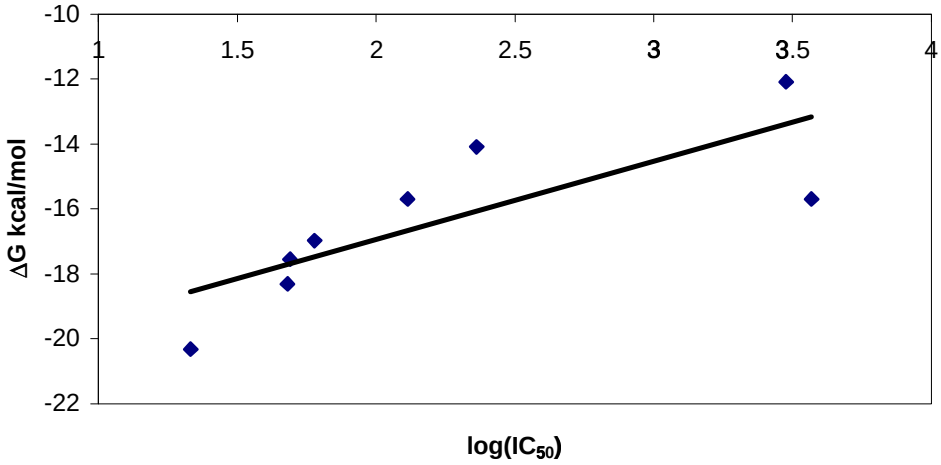
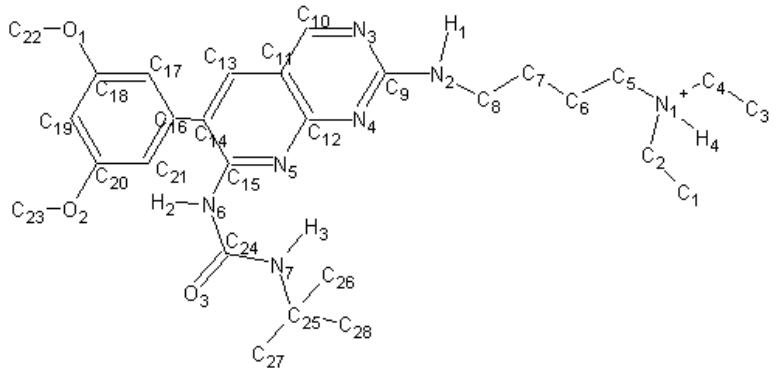
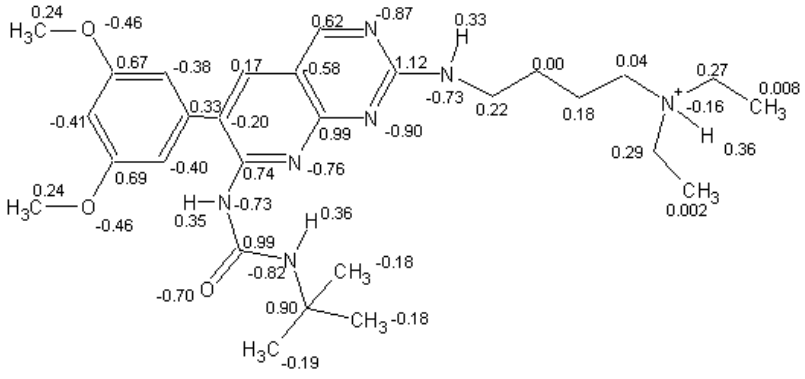


Fig. 2. Correlation between calculated binding free energy and experimental $\log(IC_{50})$ for PD173074 and 7 of its derivatives.



(a)



(b)

Fig. 3. (a) The structure of the angiogenesis inhibitor PD173074 along with its atom labels from PDB entry 2FGI.[16] (b) Atomic partial charges of the ligand atoms used in the calculations.

Table 2. Free energy derivatives measuring the significance of different charge or polar groups in affecting the binding affinity between PD173074 and FGFR1

Ligand Moiety	Charge Modifications[α]	$d\Delta G/d\lambda_i\lambda_i$ (kcal/mol)
Diethylamino group (N1)	C1: 0.002 → 0, C2: 0.29 → 0, C3: 0.008 → 0, C4: 0.27 → 0, C5: 0.04 → 0, C6: 0.18 → 0, N1: −0.16 → 0, H4: 0.36 → 0	1.57
N–H group (N2)	N2: −0.73 → −0.40, H1: 0.33 → 0	−1.25
N–H group (N6)	N6: −0.73 → −0.40, H2: 0.35 → 0.02	−0.38
N–H group (N7)	N7: −0.82 → −0.49, H3: 0.36 → 0.03	0.03
Aromatic nitrogen (N3)	N3: −0.87 → −0.11	−10.08
Aromatic nitrogen (N4)	N4: −0.90 → −0.14	−4.29
Aromatic nitrogen (N5)	N5: −0.76 → 0	−3.05
Carbonyl group (O3)	O3: −0.70 → 0, C24: 0.99 → 0.29	−0.01
Ether oxygen (O1)	O1: −0.46 → 0	−2.83
Ether oxygen (O2)	O2: −0.46 → 0	−2.20
C22	C22: 0.24 → 0	−0.06
C23	C23: 0.24 → 0	0.00
C18	C18: 0.67 → 0	−2.00
C20	C20: 0.69 → 0.02	−2.09
C9	C9: 1.12 → 0	−5.10
C10	C10: 0.62 → 0	−0.23
C12	C12: 0.99 → 0	−1.86
C15	C15: 0.74 → 0	−0.81
C8	C8: 0.22 → 0	0.09
C25	C25: 0.90 → 0	0.13

[α]Charges that were changed in calculating the free energy derivatives $d\Delta G/d\lambda_i\lambda_i$ where λ_i represents a collective charge or dipole moment. See text for details.

Table 2 shows the charge and dipole moment modifications that were carried out for functional groups of the ligand, and the corresponding free energy derivatives that gauge the significance of these charges and dipole moments in affecting binding. "Double mutations" involving the functional groups of the ligand and residues of the protein within 10 Å of these groups were also carried out. The most significant ones are shown in Table 3. The results in Table 2 show that the charge on N3 is quite significant for binding. This is consistent with the previous proposal that many small-molecule protein kinase inhibitors utilize two hydrogen bonds to recognize the linker region connecting the N-terminal and C-terminal lobes of the catalytic domain.[30] One of them is a hydrogen bond donor, the other a hydrogen bond acceptor. N3 here serves as the hydrogen bond acceptor to the backbone NH group of A564. This is further supported by the "double mutation" calculation that gave a favorable interaction of −3.87 kcal/mol (Table 3). N3 also has a useful interaction with the NH group of G567 with an interaction energy of −0.31 kcal/mol. The interactions between N3 and the sidechain guanidium group of R627, and the ammonium groups of K638 and K482 were also found to be significant (Table 3). Table 3 also shows that there are a few unfavorable protein-ligand interactions involving N3 but they are too weak to counteract the favorable interactions. Although there are a number of interactions that N3 is involved with, its hydrogen bond with the linker region appears to provide the most significant contributions to binding. The NH group N2–H1 provides the other hydrogen bond for recognizing the linker. Here, it serves as a hydrogen bond donor to the carbonyl group of A564 and "double mutagenesis" study confirms this interaction to be quite favorable (−1.69 kcal/mol). It will be useful to keep this hydrogen bond donor and acceptor pair during a lead optimization process.

Several other nitrogens also play a useful role in binding, e.g. N4. This nitrogen does not make a hydrogen bond to any residue of the protein according to the hydrogen bond visualizer in Sybyl.[31] The "double mutation" study, however, shows that N4 has an interaction with the NH group of A564 that enhances binding. An interaction with the NH group of G567 was also found but this interaction is not as

Table 3. "Double mutation" calculations that measure the strength of interactions between two functional groups[α].

Ligand Moiety	Protein Residue	Charge Modifications of Protein Residue	$d^2\Delta G/d\lambda_i d\lambda_j \lambda_i \lambda_j$ (kcal/mol)
N1	K482	C190: 0.25 → 0, N191: −0.3 → 0, H192: 0.35 → 0, H193: 0.35 → 0, H194: 0.35 → 0	1.29
N1	E571	C929: 0.14 → 0, O930: −0.57 → 0, O931: −0.57 → 0	−1.22
N1	S565	C867: 0.6 → 0.05, O868: −0.55 → 0	−0.34
N1	R570	N913: −0.4 → 0, H914: 0.3 → 0, C915: 0.5 → 0, N916: −0.45 → 0, H917: 0.35 → 0, H918: 0.35 → 0, N919: −0.45 → 0, H920: 0.35 → 0, H921: 0.35 → 0	0.55
N1	R576	N981: −0.4 → 0, H982: 0.3 → 0, C983: 0.5 → 0, N984: −0.45 → 0, H985: 0.35 → 0, H986: 0.35 → 0, N987: −0.45 → 0, H988: 0.35 → 0, H989: 0.35 → 0	0.62
N1	V631	C1396: 0.6 → 0.05, O1397: −0.55 → 0	−0.10
N2	A564	O860: −0.55 → 0, C859: 0.60 → 0.05	−1.69
N2	Y563	C853: 0.6 → 0.05, O854: −0.55 → 0	0.13
N6	K514	N400: −0.3 → 0, H401: 0.35 → 0, H402: 0.35 → 0, H403: 0.35 → 0	0.27
N6	D641	C1486: 0.14 → 0, O1487: −0.57 → 0, O1488: −0.57 → 0	−0.26
N3	A564	N855: −0.4 → −0.15, H856: 0.25 → 0	−3.87
N3	Y563	N841: −0.4 → −0.15, H842: 0.25 → 0	0.45
N3	Y563	O851: −0.65 → −0.25, H852: 0.4 → 0	0.26
N3	S565	N861: −0.4 → −0.15, H862: 0.25 → 0	0.47
N3	G567	N882: −0.4 → −0.15, H883: 0.25 → 0	−0.31
N3	V631	N1390: −0.4 → −0.15, H1391: 0.25 → 0	−0.19
N3	V493	C244: 0.6 → 0.05, O245: −0.55 → 0	0.27
N3	R627	N1351: −0.4 → 0, H1352: 0.3 → 0, C1353: 0.5 → 0, N1354: −0.45 → 0, H1355: 0.35 → 0, H1356: 0.35 → 0, N1357: −0.45 → 0, H1358: 0.35 → 0, H1359: 0.35 → 0	−0.35
N3	K638	C1460: 0.25 → 0, N1461: −0.3 → 0, H1462: 0.35 → 0, H1463: 0.35 → 0, H1464: 0.35 → 0	−0.84
N3	K482	C190: 0.25 → 0, N191: −0.3 → 0, H192: 0.35 → 0, H193: 0.35 → 0, H194: 0.35 → 0	−0.82
N4	A564	N855: −0.4 → −0.15, H856: 0.25 → 0	−0.53
N4	G567	N882: −0.4 → −0.15, H883: 0.25 → 0	−0.20
N5, N7	—	Intramolecular interaction. See Table 2 for charge modifications.	−0.24
N5	A564	N855: −0.4 → −0.15, H856: 0.25 → 0	−0.16
O3	K514	N400: −0.3 → 0, H401: 0.35 → 0, H402: 0.35 → 0, H403: 0.35 → 0	−0.21
O3	D641	C1486: 0.14 → 0, O1487: −0.57 → 0, O1488: −0.57 → 0	0.17
O1	D641	N1482: −0.4 → −0.15, H1483: 0.25 → 0	−1.76
O1	M535	S587: −0.12 → 0	0.34
O1	F642	N1491: −0.4 → −0.15, H1492: 0.25 → 0	−0.22
O2	M535	S587: −0.12 → 0	0.26
O2	K514	N400: −0.3 → 0, H401: 0.35 → 0, H402: 0.35 → 0, H403: 0.35 → 0	−1.30
C9	A564	O860: −0.55 → 0, C859: 0.60 → 0.05	−2.55
C9	E562	C839: 0.6 → 0.05, O840: −0.55 → 0	−0.57
C9	Y563	C853: 0.6 → 0.05, O854: −0.55 → 0	1.28
C9	Y563	O851: −0.65 → −0.25, H852: 0.4 → 0	−0.28
C9	A564	N855: −0.4 → −0.15, H856: 0.25 → 0	1.62
C9	S565	C867: 0.6 → 0.05, O868: −0.55 → 0	0.41
C9	V631	C1396: 0.6 → 0.05, O1397: −0.55 → 0	−0.53

[α]The charge modifications of the functional groups of the ligand are the same as in Table 2.

favorable as the interaction with A564. It may be useful to keep N4 during lead optimization. On the other hand, N5 only has a slight contribution to binding. It forms an intramolecular hydrogen to H3. According to the "double mutagensis" calculation, this interaction is somewhat enhanced, by −0.24 kcal/mol, upon binding due to the alternation of solvent-mediated interactions in going from the free ligand form to the

protein-bound form in which the solvent exposure is altered. N5 also forms a weak hydrogen bond with the NH group of A564, having a calculated pairwise interaction energy of −0.16 kcal/mol. But these interactions are too weak to make N5 a very useful contributor to binding. There are two NH groups in the urea moiety. The polarity of one of them (N7–H) hurts binding slightly. Hence, replacing this group by a nonpolar one might be useful. The polarity of the other NH group (N6–H2) closer to the pyrido[2,3-d]pyrimidine ring enhances binding modestly. The "double mutation" study shows that this NH group has a favorable interaction with the COO^- group of D641 but this interaction is counteracted by the unfavorable interactions with the ammonium group of K514. The polarity of the carbonyl group (C24–O3) of the urea moiety is also not very useful, giving a negligible free energy derivative of −0.01 kcal/mol. Since the polarity of these parts of the inhibitor are not yet doing much in binding, it will be useful to try different functional groups to see whether binding affinity can be improved.

The "single mutation" study indicates that it is unfavorable to have a positively charged ammonium group in R1 (associated with atom N1), as indicated by a positive free energy derivate of 1.57 kcal/mol. There are a number of unfavorable pairwise interactions between this ammonium group and several nearby protein functional groups. There is also a desolvation penalty of about 0.4 kcal/mol for this charged group upon binding. We have checked that this ammonium group really prefers to be in the charged form in the protein. Using the thermodynamic cycle described in Methods, the estimated free energy of protonation of the diethylamino group is 20.3 kcal/mol ($= \Delta G2 + \Delta G3 - \Delta G1 = 14.64 - 7691.4 + 7697.04 =$ 20.3 kcal/mol). This implies that the diethylamino group prefers to be charged when the ligand is bound to the protein. "Double mutation" calculations found favorable interactions of this ammonium group with the negatively charged COO^- group of E571, and the carbonyl groups of S565 and V631. However, unfavorable interactions with the $NH3^+$ group of K482, and the guanidium groups of R570 and R576 overwhelm these favorable interactions. Coupled with the unfavorable desolvation energy of the ammonium group upon binding, it might be useful to put an uncharged group here and introduce a charged group elsewhere in the inhibitor to maintain aqueous solubility and if possible improve binding affinity as well.

Consistent with experimental study,[32] the two methoxyl groups are both useful for binding. The sensitivity analysis provides further insights into how these groups achieve binding affinity. O1, O2, C18 and C20 all show favorable free energy derivatives. O1 makes a hydrogen bond to the NH group of D641, but O2 makes no hydrogen bonds to the protein, according to the hydrogen bond visualizer in Sybyl.[31] The comparable importance of O1 and O2 from the "single mutation" study suggests that O2 might have longer-range interactions with residues of the protein that make it useful for binding. The "double mutation" study indicates that O2 has a favorable interaction (−1.3 kcal/mol) with the NH_3^+ group of K514. On the other hand, the methyl groups of both methoxyl moieties show negligible $d\Delta G/d\lambda_i\lambda_i$, suggesting there is little benefit of replacing them with charged or polar groups of comparable size.

According to the results in Table 2, C9 is also important for binding. The "double mutation" study indicates that C9 has favorable electrostatic interactions with the carbonyl groups of A564, V631 and E562. The most important of these interactions is the one with A564. C9 also interacts favorably with the OH group of Y563. It is therefore useful to keep this carbon during a lead optimization process.

3.3. *Analysis of variability of amino acid types at the protein-ligand interface*

Before further evaluating the results from sensitivity analysis and using them to guide the design of new derivatives, we wanted to gain some insights into how the PD173074 may be modified to achieve specificity. We therefore analyzed the variation of amino acid types at different sites near the protein-ligand interface. Table 4 shows the results for residues that line the protein-ligand interface and have side-chains close to the binding pocket. Some sites — such as positions 512, 531, 641 and 514 — are largely conserved. It might not be useful to target such sites to achieve selectivity. Among the sites that are more variable, some have their side-chains positioned in ways that are more easily targeted by derivatives of PD173074. These include M535, I545, V559, V561, and A640 that are near the dimethoxylphenyl group

Table 4. Distribution of amino acid types at sites near the protein-ligand interface.

	L484	A640	A512	L630	E531	I545	V559	V561	A640
G	0	47	0	0	1	0	1	1	47
A	1	124	359	0	1	5	1	0	124
V	31	11	18	7	1	198	34	14	11
L	215	3	2	315	0	35	150	65	3
I	131	33	2	5	0	96	109	9	33
S	1	59	0	1	0	2	0	5	59
T	0	53	0	0	0	18	1	80	53
D	0	0	1	0	2	1	0	0	0
N	1	0	0	0	0	0	0	1	0
K	3	0	0	0	0	2	1	0	0
E	0	0	0	0	376	1	0	1	0
Q	0	0	0	0	2	2	1	11	0
R	0	0	0	0	1	0	1	1	0
H	0	1	0	0	1	1	0	0	1
F	2	0	1	20	0	0	32	66	0
C	0	54	3	0	0	6	0	0	54
W	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	1	1	18	0
M	1	0	1	37	0	3	53	113	0
P	0	0	0	0	0	14	0	0	0
–	1	0	0	0	2	0	0	0	0

	A564	D641	K514	E571	S565	G567	Y563	F642	K482	G485	V493	M535
G	3	0	0	4	20	262	0	0	11	371	1	3
A	30	1	0	16	24	2	4	0	13	5	3	1
V	73	0	0	4	3	5	1	0	6	0	27	2
L	51	0	0	5	6	4	85	23	2	0	14	212
I	20	0	0	3	0	3	9	0	3	0	10	9
S	2	0	0	41	41	6	1	0	19	2	6	1
T	1	0	0	29	21	6	0	0	8	1	2	9
D	0	380	1	93	48	1	0	0	17	0	0	1
N	0	0	0	19	47	1	0	1	16	1	1	2
K	1	0	385	32	28	3	2	0	97	0	18	0
E	0	3	0	38	71	2	2	0	75	2	8	2
Q	2	0	0	30	5	2	0	0	19	2	1	1
R	0	0	1	22	10	6	2	0	86	0	24	2
H	11	0	0	12	3	0	12	0	4	1	15	13
F	0	0	0	16	1	0	81	345	1	0	53	8
C	59	0	0	0	7	2	7	0	0	0	35	6
W	0	0	0	4	0	0	7	10	1	0	26	0
Y	7	1	0	13	0	0	168	5	1	0	114	6
M	124	0	0	4	2	0	4	0	6	0	28	107
P	1	0	0	0	48	6	0	1	2	1	0	1
–	0	0	0	0	0	74	0	0	0	1	1	1

of PD173074. Therefore, it seems profitable to modify this part of the molecule to improve potency and selectivity. The sensitivity analysis discussed earlier suggests that the polarity of the oxygens of the methoxyl groups is useful for binding. It is therefore useful to keep these oxygens to maintain their binding ability. However, these two oxygens are not likely to present significant selectivity because they either interact with backbone functional groups or conserved residues. O1 achieves its binding affinity largely by interacting with the N–H group of D641 while O2 does so with the ammonium group of K514 that is conserved among protein kinases. Therefore, a useful optimization strategy may be to replace the methyl groups of the dimethoxylphenyl moieties with other nonpolar groups to search for compounds that may improve selectivity and potency.

Site 640 is also fairly variable and appears reachable by functional groups attached to C17, which is ortho to the pyrido[2,3-d]pyrimidine ring. It may thus

be worthwhile to put different functional groups at C17 to see whether selectivity can be improved. In FGFR1, this site is occupied by a small amino acid, alanine. In other protein kinases, this site can be occupied by larger nonpolar amino acids such as valine, isoleucine and leucine, polar residues such as serine and threonine, and the small residue glycine. There are already experimental evidences that placing methyl, chloro, and bromo groups at this site can enhance binding affinity.[32]

3.4. *Derivatives of PD173074*

The insights obtained from sensitivity analysis and database analysis were further evaluated and extended by calculating the binding affinity of different derivatives of P173074 to FGFR1 (Table 5). Since sensitivity analysis suggests that the positively charged ammonium group associated with N1 is not useful for binding, we changed N1 into a carbon atom to yield ligand **9**. Indeed, removing the positive charge improved binding affinity by 1.7 kcal/mol. Since the sensitivity analysis shows that the polarity of N6–H2 in R1 is somewhat useful, we replaced R1 with a hydrocarbon chain in addition to changing N1 into a carbon atom (ligand **10**). This compound indeed binds weaker than ligand **9**. Ligand **11** further demonstrates the damaging effects of the positively charged ammonium group when this group was introduced back into ligand **10**.

As mentioned earlier, we found that introducing charges into the methyl groups of the methoxyl substituents had little effect on binding. On the other hand, the methyl groups may be useful for gaining hydrophobic interactions with the proteins. To check this, we replaced one or more methoxyl groups with OH groups (ligands **12**, **13** and **14**). All these modifications diminished binding, suggesting that having hydrophobic groups in this area is useful for binding. As mentioned earlier, this area is lined with residues that are variable among protein kinases. Hence, replacing the methyl groups of the methoxyl moieties by hydrophobic groups of varying sizes may also fine-tune binding selectivity.

The results from the sensitivity analysis indicate that the polarity of the polar groups of R2 only has modest effects on binding. We have therefore tried different R2 groups with varying length and functional groups to see whether binding affinity can be improved (Table 6). One modification in which the C=O group

Table 5. Calculated binding energy for derivatives of PD173074 (in kcal/mol).

Ligand	X	R_1	R_2	$\Delta\Delta G_{\text{binding}}$
9	3′,5′-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	NHCONHtBu	−1.73
10	3′,5′-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	$(CH_2)_3$tBu	−0.87
11	3′,5′-$(OMe)_2$	$NH(CH_2)_4NEt_2$	$(CH_2)_3$tBu	0.40
12	3′-OMe, 5′-OH	$NH(CH_2)_4NEt_2$	NHCONHtBu	1.31
13	3′-OH, 5′-OMe	$NH(CH_2)_4NEt_2$	NHCONHtBu	2.70
14	3′,5′-$(OH)_2$	$NH(CH_2)_4NEt_2$	NHCONHtBu	3.79

Table 6. Calculated binding energy for derivatives of PD173074 (in kcal/mol).

Ligand	X	R_1	R_2	$\Delta\Delta G_{\text{binding}}$
15	3′,5′-$(OMe)_2$	$NH(CH_2)_4NEt_2$	$NHCH(COCH_3)NHtBu$	−1.11
16	3′,5′-$(OMe)_2$	$NH(CH_2)_4CHMe_2$	NHCONHtBu	−0.71
17	3′,5′-$(OMe)_2$	$NH(CH_2)_4Me$	NHCONHtBu	−0.45
18	3′,5′-$(OMe)_2$	$NH(CH_2)_4Cl$	NHCONHtBu	−0.11
19	3′,5′-$(OMe)_2$	$NH(CH_2)_4NH_2$	NHCONHtBu	0.10
20	3′,5′-$(OMe)_2$	$NH(CH_2)_4NEt_2$	NHCH(OH)NHtBu	0.12
21	3′,5′-$(OMe)_2$	$NH(CH_2)_4N(Me)Et$	NHCONHtBu	0.24
22	3′,5′-$(OMe)_2$	$NH(CH_2)_4COOH$	NHCONHtBu	0.26
23	3′,5′-$(OMe)_2$	$NH(CH_2)_4NEt_2$	CH(OH)CONHtBu	0.38
24	3′,5′-$(OMe)_2$	$NH(CH_2)_4NEt_2$	$CH(NH_2)CONHtBu$	0.47
25	3′,5′-$(OMe)_2$	$NH(CH_2)_4NMe_2$	NHCONHtBu	0.93
26	3′,5′-$(OMe)_2$	$NH(CH_2)_4COO$	NHCONHtBu	3.36

was replaced by $CHCOCH_3$ to push out the C=O group was found to improve binding (Compound **15**). Replacing the N–H group closer to the aromatic ring with an aliphatic alcohol (Compound **23**) or pushing out this N–H group by inserting a hydrocarbon between it and the aromatic ring (Compound **24**) both diminish binding. This is consistent with the sensitivity analysis study that shows that the polarity of this NH group is somewhat useful for binding. Sensitivity analysis reflects that the polarity of the carbonyl group in the urea moiety has little effect on binding. Replacing this group by an O–H group has a similar effect (Compound **20**). Table 6 also contains compounds that explore the effects of different R1 groups. Replacing the damaging positively charged ammonium group by the negatively charged carboxylate group does not improve binding (Compound **22** and **26**). This suggests that this site does not favor either positively or negatively charged groups. This is probably caused by the nonpolar nature of this part of the binding site that penalizes charged groups because of desolvation penalty. On the other hand, the hydrophobic groups in R1 do seem useful as removing some of these groups diminishes binding (Compounds **16**, **17**, **18**, **19**, **21**, and **25**).

Table 7 contains results for compounds combining and extending some of the earlier modifications. Combining the replacement of the C=O group by $CHCOCH_3$ and the replacement of the protonated nitrogen with a carbon atom improves the binding affinity further and the result is almost additive (ligand 27). The binding affinity is also improved when C7 is replaced by a NH group, in addition to changing N1 to a carbon atom (ligand 28). Both ligands 27 and 28 bind stronger than ligand 9, where only the protonated nitrogen is changed. The NH group that is introduced in ligand 28 might interact with Y563, which is at a variable site, according to Table 4. Hence, this modification might improve selectivity as well as binding affinity. Introduction of the NH group at the C6 position does not help, as the resulting ligand, ligand 29, binds weaker than ligand 9. Since K482 is a nonconserved residue, we tried to replace the ethyl groups in R_1 with CH_2OH to see if one of the OH groups could interact with K482 (ligand 32). Ligand 32 binds weaker than ligand 9 suggesting that these changes are not useful. However, replacing one or both of the ethyl groups by hydrogen bond acceptor groups might improve selectivity. Ligand 30 combines this change with those done to achieve ligand 28 to achieve a slight improvement to binding. Replacing the ethyl groups with methoxyl groups (ligand 33) were found not to improve binding affinity.

When R_2 is changed to a pure hydrocarbon chain, except for the NH group at N6 (ligand 36), the binding affinity is more favorable than ligand 11 that also has this NH group changed to a hydrocarbon. This further demonstrates the positive effect of this NH group on binding. Comparing ligands 31 and 10 also supports this finding. We have also explored

Table 7. Calculated binding energy for derivatives of PD173074 (in kcal/mol).

Ligand	X	R_1	R_2	$\Delta\Delta G_{\text{binding}}$
27	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	$NHCH(COCH_3)NHtBu$	−2.81
28	$3',5'$-$(OMe)_2$	$NHCH_2NH(CH_2)_2CHEt_2$	NHCONHtBu	−1.98
29	$3',5'$-$(OMe)_2$	$NH(CH_2)_2NHCH_2CHEt_2$	NHCONHtBu	−1.32
30	$3',5'$-$(OMe)_2$	$NHCH_2NH(CH_2)_2CH(CH_2OH)_2$	NHCONHtBu	−1.16
31	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	$NH(CH_2)_2tBu$	−1.15
32	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CH(CH_2OH)_2$	NHCONHtBu	−1.06
33	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CH(OMe)_2$	NHCONHtBu	−0.89
34	$3',5'$-$(OMe)_2$	$NH(CH_2)_4NEt_2$	NHCH(COO)NHtBu	−0.51
35	$3',5'$-$(OEt)_2$	$NH(CH_2)_4NEt_2$	NHCONHtBu	−0.34
36	$3',5'$-$(OMe)_2$	$NH(CH_2)_4NEt_2$	$NH(CH_2)_2tBu$	−0.27
37	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	$NHCONH(CH_2)_4COO$	−0.06
38	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	$NHCONH(CH_2)_4NH_3$	0.04
39	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	$NHCONH(CH_2)_3COO$	0.11
40	$3',5'$-$(OMe)_2$	$NH(CH_2)_4CHEt_2$	$NHCONH(CH_2)_3NH_3$	0.52
41	$2'$-Me, $5'$-OMe	$NH(CH_2)_4NEt_2$	NHCONHtBu	0.64
42	$2',5'$-$(OMe)_2$	$NH(CH_2)_4NEt_2$	NHCONHtBu	1.08
43	$3',5'$-$(OMe)_2$	$NH(CH_2)_4NEt_2$	CH(NH3)CONHtBu	1.37

replacing this NH group by a $CHNH_3^+$ group but this modification decrease binding affinity (ligand 43).

Replacing the carbonyl group in R_2 with a negatively charged group like $CHCOO^-$ improves binding (ligand 34) but this modification is not as effective as the replacement by the $CHCOCH_3$ group demonstrated earlier by ligand 15.

In an attempt to achieve hydrogen bond interactions with the non-conserved residues R570, R627 and E571, charged groups were introduced into R_2 (ligands 37–40). Our results indicate that these changes do not improve binding, probably because these residues are quite far away from the modified groups (about 10 Å).

Since there are several non-conserved residues containing hydrophobic groups near the dimethoxylphenyl group of PD173074, we replaced the methoxyl groups with ethoxyl (ligand 35) and found a slight improvement of binding. However, a previous study showed that this modification has a negative impact on activity.[32] In ligand 42, the 3′ methoxyl-group is moved to the 2-position, whereas in ligand 41 a methyl group is placed in the 2-position instead of 3′. Both modifications lead to a decrease in binding affinity.

In summary, modifications that lead to improvement of binding affinity include replacing the positively charged ammonium group with a non-charged hydrocarbon, substituting the C=O group in R_2 with $CHCOCH_3$, and changing several hydrocarbons of R1 with polar N–H groups. Modifying the dimethoxylphenyl ring may also yield more potent and selective compounds.

4. Conclusions

Using sensitivity analysis to dissect the interactions between PD173074 and the catalytic domain of FGFR1 has identified parts of the inhibitor that are profitable to keep and parts that are useful to modify during a lead optimization process. The portion of the inhibitor that is already useful for recognizing the linker region of the protein kinase is worthwhile to keep. The positively charged diethylammonium group was found damaging to binding affinity. Unless it is important to use this group to improve aqueous solubility, it may be worthwhile to replace this positively charged ammonium group with a hydrophobic group. Our analysis suggests that the dimethoxylphenyl ring can be modified to fine tune its binding affinity. It seems profitable to keep the oxygens of the methoxyl groups but explore replacing the methyl groups with other nonpolar groups. A comparative database analysis of almost 400 protein kinases also shows that selectivity may be achieved by modifying this part of the molecule. The database analysis also suggests that introducing functional groups ortho to the pyrido[2,3-d]pyrimidine ring may improve binding selectivity. Some compounds of this type has already been tested and found to be potent towards FGFR1.[32] In the urea moiety, the polarity of the N–H group closest to the pyrido[2,3-d]pyrimidine appears to be more important to keep. Replacing the carbonyl with a $CH(COCH_3)$ group may improve binding affinity.

Acknowledgments

This work has been supported in part by the NIH, NSF, Howard Hughes Medical Institute, Accelrys, Inc., National Biomedical Computation Resource, Center for Theoretical Biological Physics at UCSD, the W. M. Keck Foundation, and the Norwegian Research Council.

References

1. D. Hanahan and J. Folkman, *Cell* **86**, 353 (1996).
2. R. Kumar and I. J. Fidler, *In Vivo* **12**, 27 (1998).
3. M.S. Pepper, *Vasc. Med.* **1**, 259 (1996).
4. R.A. Kuiper, J.H. Schellens, G.H. Blijham, J.H. Beijnen and E.E. Voest, *Pharmacol. Res.* **37**, 1 (1998).
5. Z. Szekanecz, G. Szegedi and A.E. Koch, *J. Invest. Med.* **46**, 27 (1998).
6. M.J. Tolentino and A.P. Adamis, *Int. Ophthalmol. Clin.* **38**, 77 (1998).
7. M. Klagsbrun and E.R. Edelman, *Arteriosclerosis* **9**, 269 (1989).
8. M. Klagsbrun and P.A. D'Amore, *Ann. Rev. Physiol.* **53**, 217 (1991).
9. N.S. Gray, L. Wodicka, A.M. Thunnissen, T.C. Norman, S. Kwon, F.H. Espinoza, D.O. Morgan, G. Barnes, S. LeClerc, L. Meijer, S.H. Kim, D.J. Lockhart and P.G. Schultz, *Science* **281**, 533 (1998).
10. J. Folkman and P.A. D'Amore, *Cell* **87**, 1153 (1996).
11. M. Mohammadi, S. Froum, J.M. Hamby, M.C. Schroeder, R.L. Panek, G.H. Lu, A.V. Eliseenkova, D. Green, J. Schlessinger and S.R. Hubbard, *EMBO. J.* **17**, 5896 (1998).
12. S.R. Hubbard and J.H. Till, *Ann. Rev. Biochem.* **69**, 373 (2000).
13. C.F. Wong, T. Thacher and H. Rabitz, in *Reviews*

in Computational Chemistry, Vol. 12, eds. K.B. Lipkowitz and D. B. Boyd (Wiley-VCH, New York, 1998), pp. 281.
14. C.F. Wong, P.H. Hünenberger, P. Akamine, N. Narayana, T. Diller, J.A. McCammon, S. Taylor and N.H. Xuong, *J. Med. Chem.* **44**, 1530 (2001).
15. C. Gould and C.F. Wong, *Pharmacol. Therapeutics* **93**, 169 (2002).
16. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, *Nucleic Acids Res.* **28**, 235 (2000).
17. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.* **4**, 187 (1983).
18. *Quanta*, version 2000 (Accelrys Inc., San Diego, 2000).
19. B.H. Besler, K.M. Merz and P.A. Kollman, *J. Comput. Chem.* **11**, 431 (1990).
20. U.C. Singh and P.A. Kollman, *J. Comput. Chem.* **5**, 129 (1984).
21. M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pornelli, C. Adamo, S. Clifford, J. Ochterski, J.B. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, A. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B.G. Johnson, W. Chen, M.W. Wong, J.L. Andres, M. Head-Gordon, E.S. Replogle and J.A. Pople, *Gaussian* (Gaussian Inc., Pittsburgh, 1998).
22. *CHARMm22* (Accelrys Inc., San Diego, 1992).
23. J.D. Madura, J.M. Briggs, R.C. Wade, M.E. Davis, B.A. Luty, A. Ilin, J. Antosiewicz, M.K. Gilson, B. Bagheri, L.R. Scott and J.A. McCammon, *Comput. Phys. Commun.* **91**, 57 (1995).
24. M.E. Davis, J.D. Madura, B.A. Luty and J.A. McCammon, *Comput. Phys. Commun.* **62**, 187 (1991).
25. *Handbook of Chemistry and Physics*, Vol. 56, ed. R. C. Weast (The Chemical Rubber Publishing Company, Cleveland, Ohio, 1975).
26. T.A. Halgren, *J. Comput. Chem.* **17**, 490 (1996).
27. *Molecular Operating Environment*, version 2001.01 (Chemical Computing Group, Inc., Montreal, 2001).
28. S. Hanks and A.M. Quinn, *Methods in Enzymology* **200**, 38 (1991).
29. C.I. Bayly, P. Cieplak, W.D. Cornell and P.A. Kollman, *J. Phys. Chem.* **97**, 10269 (1993).
30. P.M. Traxler, *Exp. Opin. Ther. Patents* **7**, 571 (1997).
31. *Sybyl*, version 6.7.2 (Tripos Inc., St. Louis, Missouri).
32. C.J.C. Connolly, J.M. Hamby, M.C. Schroeder, M. Barvian, G. H. Lu, R.L. Panek, A. Amar, C. Shen, A.J. Kraker, D.W. Fry, W.D. Klohs and A.M. Doherty, *Bioorg. Med. Chem. Lett.* **7**, 2415 (1997).
33. J.M. Hamby, C.J. Connolly, M.C. Schroeder, R.T. Winters, H.D. Showalter, R.L. Panek, T.C. Major, B. Olsewski, M.J. Ryan, T. Dahring, G.H. Lu, J. Keiser, A. Amar, C. Shen, A.J. Kraker, V. Slintak, J.M. Nelson, D.W. Fry, L. Bradford, H. Hallak and A.M. Doherty, *J. Med. Chem.* **40**, 2296 (1997).
34. M.C. Schroeder, J.M. Hamby, C.J. Connolly, P.J. Grohar, R.T. Winters, M.R. Barvian, C.W. Moore, S.L. Boushelle, S.M. Crean, A.J. Kraker, D.L. Driscoll, P.W. Vincent, W.L. Elliott, G. H. Lu, B.L. Batley, T.K. Dahring, T.C. Major, R.L. Panek, A.M. Doherty and H.D. Showalter, *J. Med. Chem.* **44**, 1915 (2001).
35. A.M. Thompson, C.J.C. Connolly, J.M. Hamby, S. Boushelle, B.G. Hartl, A.M. Amar, A.J. Kraker, D.L. Driscoll, R.W. Steinkampf, S.J. Patmore, P.W. Vincent, B.J. Roberts, W.L. Elliott, W. Klohs, W.R. Leopold, H.D.H. Showalter and W.A. Denny, *J. Med. Chem.* **43**, 4200 (2000).

Paper V

A new gaussian-based docking method suitable for use with homology modelled proteins

Kristin Tøndel[*,1], Endre Anderssen[1] and Finn Drabløs[2]

[1]*Department of Chemistry, Norwegian University of Science and Technology, Sem Sælands v. 14, N-7491 Trondheim, Norway*
[2]*Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, MTFS, N-7489 Trondheim, Norway*

Abstract

A new score function for virtual library screening based on gaussian density estimates is introduced. A description of the protein binding site is generated using gaussian property fields calculated in the same way as in Protein Alpha Shape Similarity Analysis (PASSA). Gaussian property fields are also used to describe the ligand properties. The score function maximises the overlap between the receptor and ligand hydrophilicity and lipophilicity fields, while minimising steric clashes. The score function is trained on 218 X-ray structures of protein-ligand complexes for which experimental binding affinities are available. The use of gaussian functions to describe the protein and ligand properties makes our score function especially suited for use with protein structure models made by homology modelling. The entire training set was docked using Tabu search for the geometry optimisation, and the resulting structures were compared to the ligand X-ray structures. Using our score function, 102 of the 218 ligand conformations were within 2 Å root mean square deviation (RMSD) of the X-ray structure, and 128 conformations were within 2.5 Å RMSD. For comparison, docking of the same set of compounds with MOE-Dock resulted in 120 of the ligand conformations within 2 Å RMSD of the X-ray structure. MOE-Dock used ~50 minutes per molecule, compared to ~5 minutes per molecule for our method. Hence, MOE-Dock performs better than our docking method, but we use only 10% of the computational time. Since our docking method is very fast, it is well suited for initial screening purposes.

Key Words: Computational docking, empirical score function, gaussian property fields, Protein Alpha Shape Similarity Analysis (PASSA), virtual screening.

[*] Corresponding author. Phone: +47 73 59 41 73. Fax: +47 73 59 16 76. E-mail: kristito@phys.chem.ntnu.no.

Introduction

The knowledge about genes and proteins associated with pathological states is increasing, especially following the human genome project. This has highly increased the potential of computer-aided drug design and virtual screening. A large variety of methods is available for small-molecular docking and virtual library screening. However, most methods are highly time consuming. They also have limitations such as neglection of receptor flexibility, inaccuracies in determination of partial charges and underestimation of hydrophobic effects. Docking methods typically use a search method to explore the conformational space of the ligand in the bound state, and a score function to guide the geometry search and to estimate the binding affinity for the different conformations. Search methods range from rigorous search methods such as simulated annealing to faster methods such as Tabu search [1] and genetic algorithms [2]. Since the number of available experimentally determined protein structures is not increasing at the same speed as the number of available protein sequences, homology modelling has great potential in structure-based drug design. However, homology models are too inaccurate to be used with most existing docking methods. It is therefore a need for new docking methods that are robust against small structural errors.

Many score functions exist for ranking drug candidates and prediction of binding affinities between a receptor and a ligand. Existing score functions can be divided into three main classes: force field-based methods, empirical score functions and knowledge-based methods. The use of score functions in drug design has recently been reviewed by Böhm and Stahl [3].

Force field-based scoring methods use nonbonded energies of molecular mechanics (MM) force fields to estimate the binding affinity. Hence, the free energy of binding in solution is substituted by an estimate of the gas-phase enthalpy of binding. Force field-based methods are generally time consuming. Examples of force field-based methods showing some success include the score function implemented in the AutoDock program [4] which utilises parameters from the AMBER force field [5], MM PB/SA [6] which complement the electrostatic interactions by a solvation term calculated by the Poisson-Boltzmann equation [7] and the newly developed OWFEG (one window free energy grid) method [7]. The OWFEG method is an approximation to the expensive first-principles method of free energy perturbation [9]. A molecular dynamics (MD) simulation is carried out with the ligand-free, solvated receptor site. The energetic effects of probe atoms on a regular grid are collected and averaged during the simulation. Three simulations are run with three different probes: a neutral atom, a negatively charged and a positively charged atom. The resulting three grids contain information on the score contributions of neutral, positively and negatively charged ligand atoms located in various positions of the receptor site. The advantages of the OWFEG method are the implicit consideration of entropic and solvent effects and the inclusion of some protein flexibility in the simulations.

Empirical score functions are generally faster than force field-based methods. The underlying idea is that the binding free energy can be interpreted as a weighted sum of localised interaction terms. The interaction terms typically represent hydrogen bonding terms, ionic interactions, hydrophobic interactions, binding entropy, etc. In addition, penalty functions for e.g. steric clashes can be added. The interaction terms are usually calculated using experimental 3D structures of receptor-ligand complexes, and the weights are estimated by multiple linear regression of experimental binding affinities. One disadvantage of empirical score functions is the dependency on the set of experimental structures used to train the functions. Usually, between 50 and 100 complexes are used to train the score functions, but recently it was shown that more than 100 complexes are needed for convergence [10]. Examples of empirical score functions showing some promise include PLP [11,12], ChemScore [13] and X-Score [14]. PLP uses a sum of pairwise linear potentials between ligand and protein heavy atoms with parameters dependent on interaction type. Each pair of interacting atoms is assigned one of three interaction types: donor and acceptor hydrogen

bonding, repulsive donor-donor and acceptor-acceptor interactions and generic dispersion of other contacts. The ChemScore function is a weighted sum of hydrogen-bonding terms, terms accounting for coordinate bonding between the ligand and metal ions placed in the protein binding pocket, hydrophobic effects and the number of rotors. The X-Score regression equation contains a van der Waals interaction term, a hydrogen bonding term, a term representing the hydrophobic effect and a torsional entropy penalty.

Knowledge-based score functions are derived by statistical analysis of structural data alone, without reference to experimentally determined binding affinities. They are based on the inverse formulation of the Boltzmann law. The frequency of occurrence of individual contacts is used as a measure of their energetic contribution to binding. The frequencies are compared to frequencies from a random or average distribution. A high frequency indicates an attractive interaction, while a low frequency indicates a repulsive interaction. Knowledge-based score functions include the Potential of Mean Force (PMF) score function [15,16,17] and DrugScore [18]. The PMF score function is a sum of distance-dependent interaction potentials for atom pairs, where both enthalpic and entropic effects are assumed to be included implicitly. In the DrugScore equation also solvent-accessible surface dependent singlet potentials for protein and ligand atoms are included.

Recently, a comparison of eleven score functions using the same set of experimental structures was published [19]. This study indicates that X-Score and DrugScore are the score functions most suited for use with conformational sampling, since they produce a funnel-shaped energy surface for protein-ligand complexation, and therefore will most likely lead to a faster convergence to the global minimum. This study also indicates that a combination of several different score functions might be advantageous. X-Score, DrugScore and PLP were the score functions showing most promise in this study. However, these score functions give only moderate correlation between the predicted (using the experimentally determined conformation) and the experimental binding affinities for these 100 structures. Most of the score functions tested in this study predict hydrophilic interactions better than hydrophobic interactions. Hence, this study indicates a need for a fast and more accurate method for ranking a large number of ligands according to success of binding to a receptor.

In this work we have developed a new empirical score function for estimation of binding affinities to a receptor using gaussian property distributions for both the protein and the ligands. The score function evaluates only the match between the lipophilicity and hydrophilicity of the receptor and the ligand, in addition to describing van der Waals effects. This makes it easy to interpret and robust. The fact that the score function is based on gaussian density estimates makes it more robust against the errors typically found in homology models, since gaussian functions give a less detailed representation than force field models, and they have neither steep derivatives nor singularities [20]. Hence, this score function will be well suited for virtual screening using protein structure models built by homology modelling. Since this score function is robust against small structural variations, including protein flexibility is less important than in many other docking methods. Because of the very simplified description of the protein-ligand interactions, the accuracy of our new score function can not be compared to score functions that take e.g. partial charges and electrostatics into account. However, the speed of our calculations makes this method an effective tool for exploration of the ligand conformational space and generation of starting conformations for more accurate docking methods.

The protein binding site properties are mapped using the newly developed method Protein Alpha Shape Similarity Analysis (PASSA) [21], while the ligand properties are described using gaussian property distributions similar to those used in Comparative Molecular Similarity Index Analysis (CoMSIA) [22]. Both PASSA and CoMSIA work by assigning property distributions to each atom, so that the molecules are described by the spatial distribution of their interactions. At each point of a 3D grid the values of the molecular similarity fields are computed. A molecular similarity index is based on atomic parameters such as lipophilicity, hydrogen bond donor and acceptor properties, etc.

A gaussian function with the intensity of the atomic parameter, and a standard deviation (σ) corresponding to the atomic radius is centred at each atom. For each physicochemical property, the value in a grid point is computed as the sum of the contributions from all gaussian functions representing that property (see Equation 1).

$$F(q,j) = \sum_{i=1}^{n} \frac{\omega_{ik}}{(\sigma_i \sqrt{2\pi})^3} \cdot e^{\frac{-r_{iq}^2}{2\sigma_i^2}} \qquad (1)$$

F is the value of the similarity field in grid point q of molecule j, ω_{ik} is the value of the physicochemical property k of atom i, r_{iq} is the distance between grid point q and atom i and σ_i corresponds to the atomic radius of atom i.

PASSA [21] converts the discreet information contained in the placement of geometrical objects known as alpha spheres and the positions of protein atoms to a continuous field using gaussian density estimates. An alpha sphere is a sphere that contacts four protein atoms on its surface and has no atoms contained internally. Centres of alpha spheres have been found to correspond well with the placement of atoms in bound ligands [23]. Alpha spheres are determined geometrically, using only the positions and radii of the heavy atoms. This eliminates the need for placing hydrogens and determining protonation states and partial charges. The alpha spheres are classified as hydrophobic or hydrophilic depending on the protein atoms that they contact. In PASSA, gaussian functions (as shown in Equation 1) are centred at dummy atoms placed at each alpha sphere centre, and at all protein atoms. A 3D grid is placed around the binding site of the protein, and in each grid point the sum of the contributions from all gaussian functions is computed. The use of gaussian functions with a very simple partitioning according to the hydrophilic or hydrophobic nature of the alpha spheres, reduces some of the problems associated with force field models [20].

Recently, new docking methods have been reported that utilise gaussian functions to describe protein-ligand interactions. Schafferhans and Klebe [24] published a method for computational docking of ligands into protein binding sites that is especially suited for protein structures derived by homology modelling. This method uses gaussian functions to represent the physicochemical properties of the receptor and the ligand, and incorporates ligand information into the protein structure modelling procedure. Another docking method that utilises gaussian functions is the method developed by McGann *et al*. [25] that acts as a filter to reduce the search space for other docking methods. This method only accounts for shape, and minimises steric clashes between the receptor and ligand atoms. By using gaussian functions representing hydrophilicity and lipophilicity, in addition to describing van der Waals effects, we hope to be able to describe protein-ligand interactions better than methods that only account for steric clashes.

Methods

Computation of gaussian property fields

The properties of the protein binding site were mapped in the same way as in PASSA [21]. A 3D grid centred at the ligand and extended to 3 Å outside the ligand was used to compute the gaussian property fields. A grid spacing of 0.5 Å was used. The dummy atoms placed in each alpha sphere centre had the properties of either an oxygen atom or a carbon atom, depending on whether the alpha spheres were classified as hydrophilic or hydrophobic. The gaussian functions centred at the dummy atoms were given unit weight for either the hydrophilic or the hydrophobic field, according to the properties of the alpha spheres. The standard deviation of the gaussian functions (σ in Equation 1) corresponded to half the van der Waals radius for both dummy atoms and protein atoms. To include steric effects, the gaussian functions centred in protein atoms were given the weight –1 for both fields. In addition to the hydrophilic and hydrophobic field, a separate van der Waals field was generated for the protein. Also for this field the standard deviation of the gaussian functions corresponded to 0.5 times the van der Waals radius for the protein atoms, and the gaussian functions were all given unit weight.

The ligand properties were described using gaussian property fields similar to those used in CoMSIA [22]. In the same way as for the protein, gaussian functions with unit weight and standard deviation corresponding to 0.5 times the atomic van der Waals radius were centred in each ligand atom. The atomic properties were determined using the pharmacophore functions [26] in Molecular Operating Environment (MOE) [27]. These functions return either zero or one, depending on whether the atom is of the specified pharmacophoric type or not. The following properties were used: hydrophilicity, lipophilicity, hydrogen acceptor and hydrogen donor. The van der Waals field for the ligand was computed using gaussian functions with standard deviation corresponding to 0.5 times the van der Waals radius of the ligand atoms and weight equal to the van der Waals radius.

All scripts are written in Scientific Vector Language (SVL) [27] and can be obtained from the authors upon request.

Variables used to describe the binding affinity

In each grid point, the products of the ligand and protein gaussian fields were computed. These product values were then summed over all grid points, giving one variable describing the match between the given ligand and protein fields. The variable ”protein lipophilicity * ligand lipophilicity (lip_lip)”, for example, describes the match between the protein and ligand lipophilicity fields, summed over all grid points. For the protein, only lipophilicity and hydrophilicity are considered. For the ligand, the following variables are used: lipophilicity, hydrophilicity and hydrogen donor or acceptor properties. No hydrogens or partial charges are taken into account in the calculations.

Training of the empirical score function

Five different sets of experimental structures of protein-ligand complexes for which the binding affinity ($\Delta G_{binding}$) is known were used to train the score function. One was the set of 50 complexes used by Baxter *et al*. [1], to validate a flexible docking method using Tabu search. In addition to these 50 structures, we used the 100 complexes reported in [19]. These 100 structures have been used by Wang *et al*. to evaluate the performance of eleven score functions for molecular docking [19]. The five peptide structures reported in [28], the 170 protein-ligand complexes used to train the

empirical score function SCORE [10] and the 19 complexes reported in [29] were also added to our training set. All together we used 218 different protein-ligand complexes to train our score function.

Ions and other co-factors were treated as a part of the receptor when not directly bound to the ligand. If more than one ligand molecule were present in the structure, only one of them was kept. Water molecules (and hydrogen atoms) are not considered in the calculations. Our calculations do not separate hydrogen donors from acceptors on the protein, and no directions are considered when estimating hydrophilic interactions. Since our calculations are independent of placement of hydrogen atoms, the fact that a hydrogen atom can form hydrogen bonds only with atoms pointing towards it is not accounted for. This is one of the major weaknesses of our method. We plan to account for this in the next version.

We used Partial Least Squares (PLS) regression in Unscrambler® [30] with full cross-validation to fit the regression parameters for prediction of binding affinities. The variables describing the match between the protein and ligand fields were centred and standardised (divided by the standard deviation for each variable) prior to the regression analysis.

Docking using Tabu search

The geometry search routines used are the same as those used for the Tabu searching in MOE-Dock [1,27]. Tabu search is a stochastic searching algorithm that maintains a list of previously visited conformations. These conformations are forbidden (tabu) to future moves. A new conformation is compared to the conformations in the list by calculating the root mean square deviation (RMSD) between the Cartesian coordinates of the new conformation and those of every entry in the list. If the RMSD value is below a specified value, the conformations are considered to be the same, and the move is tabu.

In the geometry search, a fast version of the score function is used. In this version, only the values of the protein fields in the positions of the ligand atoms are used, instead of all grid point values. This speeds up the computation of the score values. In addition, the ligand property fields are substituted by vectors with zero or unit entries according to the atomic properties. Hence, no gaussian functions are used for the ligand. The ligand van der Waals field is substituted with the van der Waals radii of the ligand atoms. The van der Waals term was given unit regression coefficient in the score function used for the geometry search. To further penalise steric clashes, a sigmoid function of the van der Waals term was added to this score function. The receptor atoms are kept in fixed positions during the geometry search.

To avoid steric clashes between ligand functional groups, the Lennard-Jones potential of the ligand is evaluated. A threshold value of 500 kcal/mol is used in the geometry search. Several different threshold values were tested (100, 200, 500 and 1000 kcal/mol), but changing this parameter did not have a significant effect on the results.

The structures resulting from each Tabu search run (the results from several iterations) are ranked using the same score function as used in the geometry search, and the best ranked conformation is used for the final prediction of the binding affinity. The binding affinity is predicted using a slower, more accurate score function, computed using all grid points. In this version of the score function, gaussian functions are used to describe both the protein and the ligand properties.

Method testing

To test the performance of our new empirical score function, a docking analysis was performed with all protein-ligand complexes in the training set. All calculations were done in MOE [27], using

the molecular mechanics force field MMFF94 [31]. A smooth non-bonded cut-off of 10-12 Å was used.

Ten Tabu search runs of 1000 iterations each were performed, and binding affinities were predicted using the score function that utilises the entire grid. The docking calculations were started from the X-ray structures.

For comparison, a similar docking study was carried out using MOE-Dock [1,27] with Tabu search. Prior to the docking analysis with MOE-Dock, hydrogen atoms were added to the X-ray structures, and optimised to an RMS gradient of 1 with MMFF94 and a smooth non-bonded cut-off of 10-12 Å. MOE-Dock calculates the potential energy grids only once, at the beginning of the docking procedure. Hence, protein flexibility is not taken into account in the MOE-Dock calculations.

Results and discussion

Gaussian property description

Figure 1 shows an example of a gaussian property description of a known protein-ligand complex (RCSB Protein Data Bank (PDB) [32,33] entry 1AGP). Only the hydrophilicity field for the protein is shown, together with the ligand structure. As seen from the figure, the hydrophilic groups of the ligand and the hydrophilicity field of the protein match to a high degree for this complex.

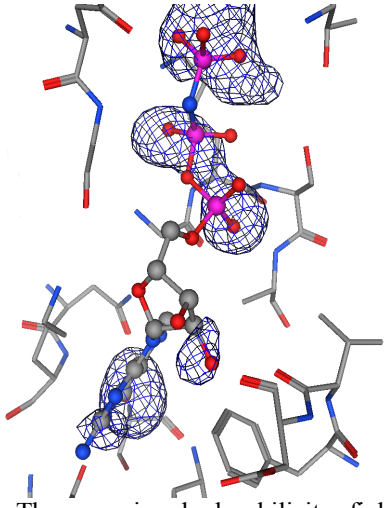


Figure 1. The gaussian hydrophilicity field for the protein in PDB entry 1AGP plotted together with the ligand. The blue mesh indicates the hydrophilicity field.

Training of the empirical score function

Two different score functions were made using the five sets of experimental structures described above, a fast score function to be used for the geometry search, and a more accurate function for the final estimation of binding affinities. The designed samples from the set made by Baxter *et al.* [1] (DFR4, TSC2 and TMT1) were excluded since they were not real X-ray structures. PDB [32,33] entries 1FKB and 2XIS were also removed from the training set due to problems with the interpretation of the connection of ligand atoms in the X-ray structure.

The variables used to score the match between the protein and ligand properties are shown in Table 1.

Table 1. The field variables used in the score functions.

Variable	Description
hyd_hyd	protein hydrophilicity * ligand hydrophilicity
lip_lip	protein lipophilicity * ligand lipophilicity
hyd_lip	protein hydrophilicity * ligand lipophilicity
lip_hyd	protein lipophilicity * ligand hydrophilicity
lip_hacc	protein lipophilicity * ligand hydrogen acceptor
vdw_vdw	protein van der Waals field * ligand van der Waals field

The fast version of the score function used for the geometry search is given in Equation 2. The correlation between the score value (from the cross-validation, keeping the van der Waals terms out) and the experimental binding affinity for this score function is 0.62. The terms of the score function accounting for van der Waals effects were kept out of the regression analysis, since the training set consists only of structures without severe steric clashes. The regression coefficient resulting from the PLS regression would not be useful in the geometry search, where steric clashes have to be accounted for. We tested the performance of the docking method using several different values for the van der Waals parameter, and for our training set, the docking method performed best when the van der Waals parameter was given unit weight.

To further penalise steric clashes, a sigmoid function was added to the van der Waals term. The parameters of this function were chosen based on observed values for the X-ray structures in the training set. Using this function leads to a steep increase in the steric clash penalty at values of vdw_vdw above 0.1, since this is the highest value of vdw_vdw observed in the training set (Figure 2 A). The steric clash penalty reaches a constant value when vdw_vdw passes 0.2. The parameter for this term (having the value 100) was chosen to give this term high weight compared to the other terms. This sigmoid function has no effect for low values of vdw_vdw. Structures for which this sigmoid function has a higher value than 90 are considered so wrongly placed that they are given a high, positive value of 1000 for the score. The steric clash penalty is shown as a function of vdw_vdw in Figure 2 B.

$$\text{Score} = -2.976\ \text{hyd\_hyd} - 9.187\ \text{lip\_lip} + 2.775\ \text{hyd\_lip} - 4.1\ \text{lip\_hyd} + 5.028\ \text{lip\_hacc} + \text{vdw\_vdw} + 100/(1+e^{(-50*\ \text{vdw\_vdw}+8)}) \quad (2)$$

The score function in Equation 2 can not be used for binding affinity prediction, because of the van der Waals terms that were added to it. This function is only suitable for finding the best ligand conformations in a geometry search. The score function using gaussian functions for both protein and ligand property fields and summation over all grid points is more accurate, and was used to estimate the binding affinities for the best conformations from each docking run. This score function is given in Equation 3. The correlation between the predicted binding affinity (from the cross-validation) and the experimental binding affinity for this score function is 0.64. Since we assume that the best conformation resulting from a docking run can be compared to an X-ray structure in the sense that they contain no severe steric clashes with the protein structure, we use the regression coefficient from the PLS regression for the van der Waals parameter in this score function. Hence, the estimate for the binding affinity produced by this score function lies in the correct range.

$$\Delta G_{binding} = -2.154\ \text{hyd\_hyd} - 8.719\ \text{lip\_lip} + 3.199\ \text{hyd\_lip} - 2.93\ \text{lip\_hyd} + 4.035\ \text{lip\_hacc} - 3.464\ \text{vdw\_vdw} \quad (3)$$

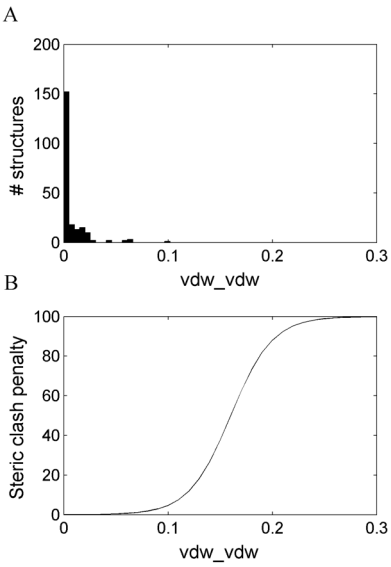


Figure 2. A: Histogram over the values of the variable vdw_vdw (protein van der Waals field * ligand van der Waals field) for the X-ray structures in the training set. B: The steric clash penalty as a function of vdw_vdw.

Method testing

To test the performance of our new empirical score function, a docking analysis was performed with all protein-ligand complexes in the training set. The same set of structures was used both for training of the score function and for testing of the docking method. However, the wide variety of structures present in the data set combined with extensive cross-validation and a relatively small number of parameters ensure that the score function has not been overfitted. The test will therefore still be valid. The root mean square deviation (RMSD) between the X-ray ligand structures and the ligand structures resulting from the docking analysis was calculated. The results are given in Figure 3. The predicted binding affinities found using the score function in Equation 3 are plotted against the experimental binding affinities in Figure 4.

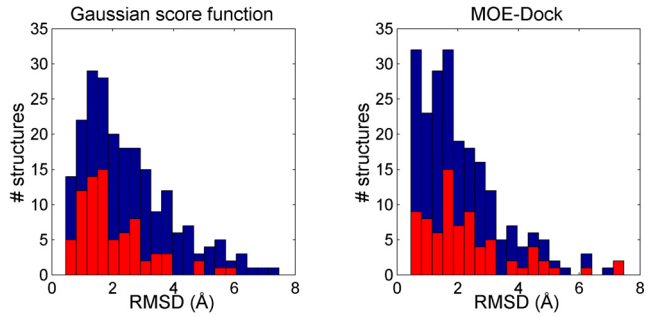


Figure 3. Left: Histogram over RMSD values between the X-ray ligand structures in the training set and the ligand structures resulting from 10 Tabu runs á 1000 iterations using our gaussian-based score function. The fraction of the complexes having experimental binding affinities below –40 kJ/mol is shown in red. This docking procedure uses ~5 minutes per molecule.
Right: Histogram over RMSD values between the X-ray ligand structures in the training set and the ligand structures resulting from 10 Tabu runs á 1000 iterations with MOE-Dock. The fraction of the complexes having experimental binding affinities below –40 kJ/mol is shown in red. This docking procedure uses ~50 minutes per molecule.

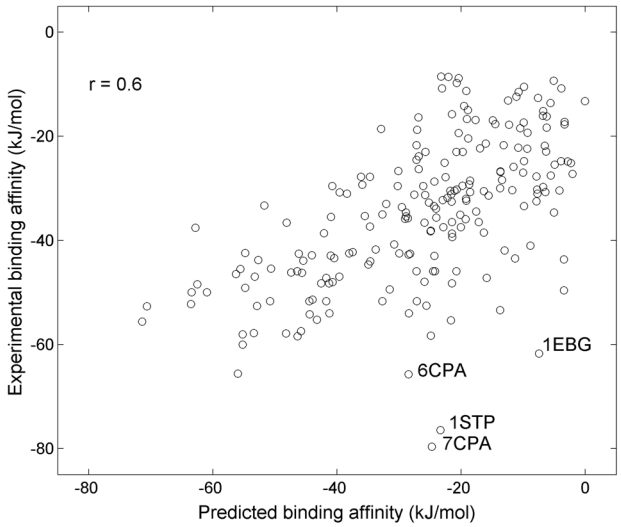


Figure 4. Predicted binding affinities (found using Equation 3) for the ligand structures resulting from the docking analysis plotted against the experimental binding affinities.

The following structures (from the PDB [32,33]) were kept out of the plot in Figure 4 because they were severe outliers: 1L83, 2TMN, 5TMN, 6TMN, 2SNS, 9RUB and 2CTC. These seven X-ray structures all contain bond lengths and angles that are not frequently observed, caused by e.g. an ion-containing ring structure. This leads to a very high value for the estimated internal energy. Figure 4 shows that 1STP, 6CPA, 7CPA and 1EBG are false negatives (shown in the lower, right part of the figure). 1EBG contains two ion bonds, and therefore binds much stronger than predicted. 1STP binds to the protein through several hydrogen bonds, and since our method is unable to fully represent hydrogen bonds, the binding affinity is underestimated. 6CPA and 7CPA contain hydrophobic groups that protrude towards the solvent. Alpha spheres can only represent ligand atoms bound in a protein cavity. Hence, interactions on the outer surface of the protein are ignored. The contribution of the protruding groups to binding is therefore not included, and the binding affinity is underestimated.

The histograms in Figure 3 show that MOE-Dock performs better than our docking method, but MOE-Dock uses ten times as much computational time (~5 minutes versus ~50 minutes per molecule). Using our score function, 102 of the 218 ligand conformations were within 2 Å RMSD of the X-ray structure and 128 conformations were within 2.5 Å RMSD, while docking with MOE-Dock resulted in 120 of the ligand conformations within 2 Å RMSD. The two histograms showing the distribution of obtained RMSD values for the two docking methods are almost identical, except for the first column, representing the number of structures within 0.6 Å RMSD of the X-ray structure. The histograms also indicate that our docking method performs better than MOE-Dock for ligands having high affinity for the receptor (shown in red). The shape of the curve in the histogram for our gaussian score function corresponds to what one might expect from a score function that is robust against small structural errors. That is, the purpose of this score function is to find a reasonable ligand conformation for a large number of protein-ligand complexes, not to find the absolutely correct conformation. The level of accuracy of our score function might not be sufficient for a stand-alone docking procedure, but since our docking method is very fast, it is well suited for generation of starting conformations for more accurate docking.

Protein flexibility is not explicitly taken into account in the docking calculations. The use of gaussian functions to describe the protein and ligand properties will partly compensate for this, since this makes the score function robust against small structural variations. Hence, protein flexibility is less critical than for many other docking methods.

Examples from the docking analysis with the gaussian-based score function

To illustrate in what cases our gaussian-based docking method is likely to succeed in predicting the binding affinity and the structure of a ligand-receptor complex, we show some examples from the docking analysis performed on our training set.

Three examples where our method succeeds in reproducing the ligand X-ray structure are PDB [32,33] entries 1E96, 1ETS and 1HVI. The RMSD values between the docked conformations and the ligand X-ray structures are 0.46 Å, 0.57 Å and 0.93 Å, respectively. Figures 5, 6 and 7 show the docking results for these three PDB entries.

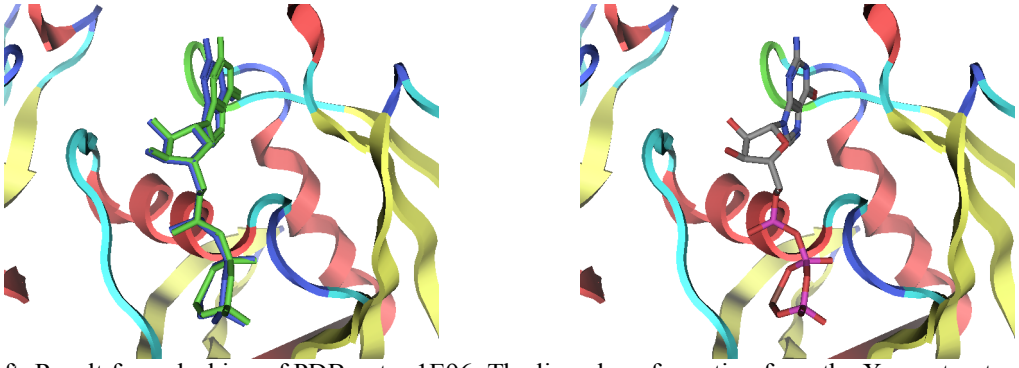


Figure 5. Left: Result from docking of PDB entry 1E96. The ligand conformation from the X-ray structure is rendered in green, while the docked conformation is rendered in blue. The RMSD value between the docked conformation and the ligand X-ray structure is 0.46 Å.
Right: The X-ray structure from PDB entry 1E96.

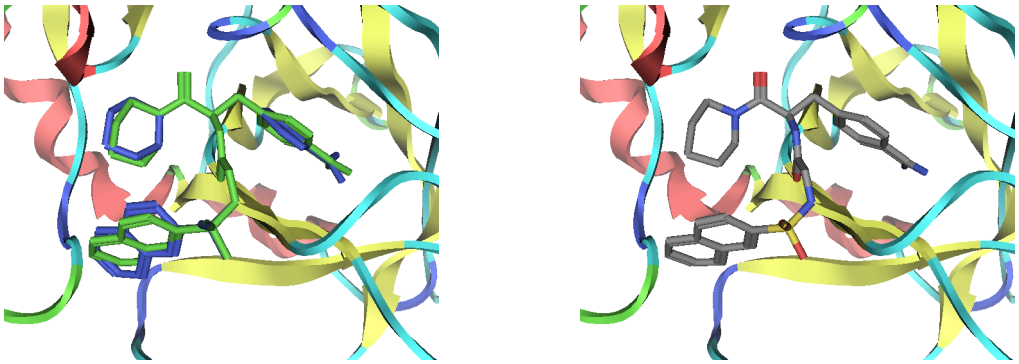


Figure 6. Left: Result from docking of PDB entry 1ETS. The ligand conformation from the X-ray structure is rendered in green, while the docked conformation is rendered in blue. The RMSD value between the docked conformation and the ligand X-ray structure is 0.57 Å.
Right: The X-ray structure from PDB entry 1ETS.

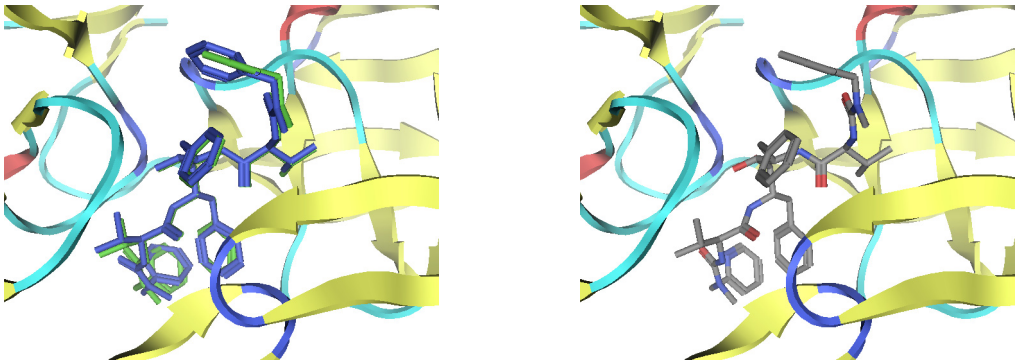


Figure 7. Left: Result from docking of PDB entry 1HVI. The ligand conformation from the X-ray structure is rendered in green, while the docked conformation is rendered in blue. The RMSD value between the docked conformation and the ligand X-ray structure is 0.93 Å.
Right: The X-ray structure from PDB entry 1HVI.

One example where our docking calculations resulted in a high RMSD value between the X-ray structure and the docked structure is PDB entry 1TET. Figure 8 shows the docked ligand conformation together with the X-ray structure of the complex.

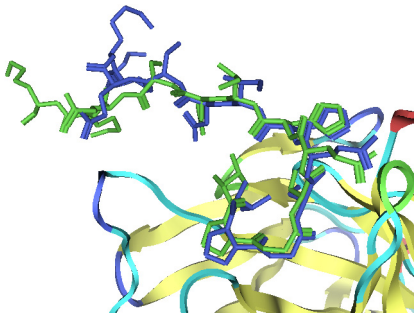


Figure 8. Result from docking of PDB entry 1TET. The ligand conformation from the X-ray structure is rendered in green, while the docked conformation is rendered in blue. The RMSD value between the docked conformation and the ligand X-ray structure is 3.70 Å.

Figure 8 shows that for 1TET we have very good match between the docked conformation and the X-ray ligand conformation in the region binding to the protein, while the part of the ligand pointing towards the solvent is flipped outwards. The reason is that our algorithm considers only cavities in the protein structure, since alpha spheres are placed in protein cavities. This leads to a very high RMSD value, even though our docking calculations succeeded for the part of the ligand that is relevant for binding to the protein.

Our docking calculations also failed to reproduce the ligand X-ray structure in PDB entry 1CPS. As shown in Figure 9 (left), the structure of the ligand in PDB entry 1CPS is flipped 180° in the docked conformation. One possible reason is that our algorithm does not represent hydrogen bonds fully. As seen from Figure 9 (right), the ligand in PDB entry 1CPS is stabilised in the bound conformation by several hydrogen bonds.

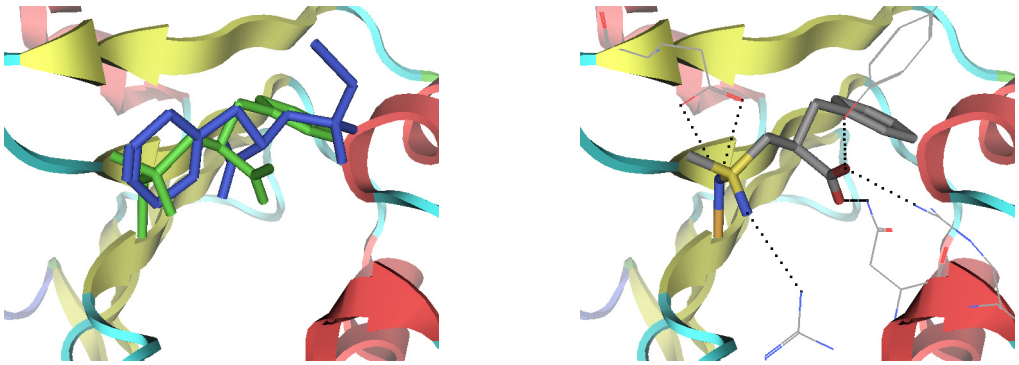


Figure 9. Left: Result from docking of PDB entry 1CPS. The ligand conformation from the X-ray structure is rendered in green, while the docked conformation is rendered in blue. The RMSD value between the docked conformation and the ligand X-ray structure is 6.18 Å.
Right: The X-ray structure from PDB entry 1CPS. Possible hydrogen bonds to the protein are shown (drawn using MOE [27]).

The result from the docking analysis of PDB entry 1DBK (Figure 10) demonstrates that our algorithm predicts hydrophobic interactions quite well. The RMSD value between the docked ligand conformation and the ligand X-ray structure is very high for 1DBK. In the same way as for 1CPS, the ligand structure is flipped 180°. However, Figure 10 shows that the ligand structure is

almost symmetric, and our algorithm succeeds in placing the hydrophobic ring structures. Since the ligand is flipped, an almost correct ligand placement leads to a very high RMSD value (5.86 Å).

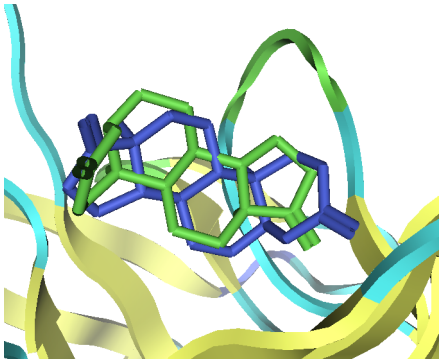


Figure 10. Result from docking of PDB entry 1DBK. The ligand conformation from the X-ray structure is rendered in green, while the docked conformation is rendered in blue. The RMSD value between the docked conformation and the ligand X-ray structure is 5.86 Å.

The results from our docking analysis show that our method predicts hydrophobic interactions better than hydrophilic interactions. One reason is that hydrophilic interactions are more direction-specific than hydrophobic interactions. The ligands in PDB [32,33] entries 1ETS and 1HVI both contain several hydrophobic groups (Figures 6 and 7). Since our method is independent of placement of hydrogen atoms and partial charges, we are not able to fully account for the formation of hydrogen bonds. Hydrogen bond formation is very dependent on the direction in which the hydrogen atom points. We are planning to account for possible hydrogen bond formation in the next version of the score function. Our method is not suitable in cases where the ligand makes an ion bond to the protein. However, our method succeeds to a high degree in placing the hydrophobic parts of the ligands. The results also show that our method works best in cases where the ligand is bound in a well-defined cavity of the protein. This is not surprising since alpha spheres only describe protein cavities, not the outer surface of the protein. Hence, our gaussian-based docking method is most likely to succeed in cases where the protein has a well-defined binding pocket and the ligand is not bound to the protein by an ion bond. The docking method presented here succeeds in reproducing the ligand conformations found in X-ray structures in most cases, and the speed of the calculations makes it a useful tool for fast drug candidate screening. The use of gaussian functions to describe the molecular properties makes this docking method suitable for use with homology modelled protein structures.

Conclusion

A new score function for virtual library screening is introduced, that use gaussian functions to describe protein-ligand interactions. This score function accounts for hydrophilicity and lipophilicity of the protein, and hydrophilicity, lipophilicity, hydrogen donor and acceptor potential for the ligand. In addition, van der Waals effects are taken into account. Neither hydrogens nor partial charges are considered. The use of gaussian functions makes this score function relatively robust against small structural errors, like those typically found in homology models. The accuracy of our score function can not be compared to that of more complex score functions that account for e.g. hydrogens, partial charges and electrostatics, but the speed of this method makes it useful for fast screening and generation of starting conformations for other docking methods. With this method we are able to dock 200-300 ligands per day (~5 minutes per molecule) on a Linux cluster. For comparison, the more accurate docking method MOE-Dock uses ~50 minutes per molecule on the same set of ligands. Using our score function, 102 of the 218 ligand conformations were within

2 Å RMSD of the X-ray structure, while docking of the same set of compounds with MOE-Dock resulted in 120 of the ligand conformations within 2 Å RMSD of the X-ray structure. Hence, MOE-Dock performs best in terms of identifying the correct conformation, but the combination of speed and reasonable accuracy makes our method more suitable for use in pre-screening.

Acknowledgement

We thank The Norwegian Research Council for financial support.

References

1. Baxter, C.A., Murray, C.W., Clark, D.E., Westhead, D.R. and Eldridge, M.D., Prot. Struct. Func. Gen., 33 (1998) 367.

2. Halperin, I., Ma, B., Wolfson, H. and Nussinov, R., Prot. Struct. Func. Gen., 47 (2002) 409.

3. Böhm, H.-J. and Stahl, M., Lipkowitz, K.B. and Boyd, D.B. (Eds.), In *Reviews in Computational Chemistry,* The Use of Scoring Functions in Drug Discovery Applications, Wiley-VCH, New York, 2002, 18, pp. 41-87.

4. Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J., J. Comput. Chem., 19 (1998) 1639.

5. Weiner, S.J., Kollman, P.A. and Case, D.A., J. Am. Chem. Soc., 106 (1984) 765.

6. Massova, I. and Kollman, P.A., J. Am. Chem. Soc., 121 (1999) 8133.

7. Nicholls, A. and Honig, B., Science, 268 (1995) 1144.

8. Pearlman, D.A. and Charifson, P.A., J. Med. Chem., 44 (2001) 502.

9. Meirovitch, H., Lipkowitz, K.B. and Boyd, D.B. (Eds.), In *Reviews in Computational Chemistry*, Calculation of the Free Energy and the Entropy of Macromolecular Systems by Computer Simulation, Wiley-VCH, New York, 1998, 11, pp. 1-74.

10. Wang, R., Liu, L., Lai, L. and Tang, Y., J. Mol. Model., 4 (1998) 379.

11. Gelhaar, D.K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.B., Fogel, L.J. and Freer, S.T., Chem. Biol., 2 (1995) 317.

12. Gelhaar, D.K., Bouzida, D. and Rejto, P.A., Parrill, L. and Reddy, M.R. (Ed.), In *Rational Drug Design: Novel Methodology and Practical Applications*, American Chemical Society, Washington, DC, 1999, pp. 292-311.

13. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., J. Comput. Aided Mol. Des., 11 (1997) 425.

14. Wang, R., Lai, L. and Wang, S., J. Comput. Aided Mol. Des., 16 (2002) 11.

15. Muegge, I. and Martin, Y.C., J. Med. Chem., 42 (1999) 791.

16. Muegge, I., Perspect. Drug Discovery Des., 20 (2000) 99.

17. Muegge, I., J. Comput. Chem., 22 (2001) 418.

18. Gohlke, H., Hendlich, M. and Klebe, G., J. Mol. Biol., 295 (2000) 337.

19. Wang, R., Lu, Y. and Wang, S., J. Med. Chem., 46 (2003) 2287.

20. Wieman, H., Tøndel, K., Anderssen, E. and Drabløs, F., Mini-Reviews in Medicinal Chemistry, Homology-based modelling of targets for rational drug design, 2003, In Press.

21. Tøndel, K., Anderssen, E. and Drabløs, F., J. Comput. Aided Mol. Des., 16 (2002) 831.

22. Klebe, G., Abraham, U. and Mietzner, T., J. Med. Chem., 37 (1994) 4130.

23. J. Liang, H. Edelsbrunner and C. Woodward., Protein Sci., 7 (1998) 1884.

24. Schafferhans, A. and Klebe, G., J. Mol. Biol., 307 (2001) 407.

25. McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A., and Brown, F. K., Biopolymers, 68 (2003) 76.

26. Sheridan, R.P. and Bush, B.L., J. Chem. Info. Comp. Sci., 33 (1993) 756.

27. Molecular Operating Environment™, Version 2002.03, Chemical Computing Group, Inc., 2002.

28. Rognan, D., Lauemøller, S.L., Holm, A., Buus, S. and Tschinke, V., J. Med. Chem., 42 (1999) 4650.

29. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.

30. The Unscrambler®, Version 7.6 SR-1, CAMO ASA, 2000.

31. Halgren, T.A., J. Comp. Chem., 17 (1996) 490.

32. The RCSB Protein Data Bank, http://www.rcsb.org/pdb/.

33. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235.

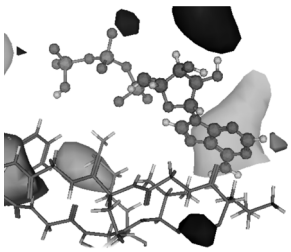
Paper VI

Design of selective inhibitors of Tyrosine kinase 2

Kristin Tøndel[*,1] and Finn Drabløs[2]

[1]*Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*
[2]*Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, MTFS, N-7489 Trondheim, Norway*

Table of Contents graphic



Abstract

Possible functional groups of a selective inhibitor of Tyrosine kinase 2 (Tyk2) have been proposed earlier by our group, based on results from Protein Alpha Shape Similarity Analysis (PASSA). The database of the National Cancer Institute (NCI) was searched for existing drugs having these functional groups. The hits from this pharmacophore search were evaluated by computational docking in a homology model of Tyk2. Additional structures having the required functional groups were created using *de novo* ligand design. The most promising structures were tested for selectivity by computational docking in seven protein kinase structures, in addition to Tyk2. The results from our docking analysis indicated that none of the structures present in the NCI database can be used to inhibit Tyk2 selectively, but five of the structures generated by *de novo* ligand design gave very promising results.

Key Words: Tyk2, tyrosine kinase, Protein Alpha Shape Similarity Analysis (PASSA), inhibitor design, selectivity, homology modelling, pharmacophore search, molecular docking.

[*] Corresponding author. Phone: +47 73 59 41 73. Fax: +47 73 59 16 76. E-mail: kristito@phys.chem.ntnu.no.

Introduction

Protein kinases contribute to regulation and coordination of e.g. metabolism, gene expression, cell growth, cell motility, cell differentiation and cell division.[1] The Janus kinase (Jak) family of non-receptor tyrosine kinases consists of four known mammalian proteins (Tyk2, Jak1, Jak2 and Jak3) that play a critical role in initiating signalling cascades of a large number of cytokine receptors.[2, 3, 4, 5] All Jak family kinases possess a carboxyl-terminal tyrosine kinase catalytic domain, a central kinase-like domain, and a large amino-terminal region, which has been subdivided into five Jak homology regions (JH7 to JH3) based on sequence conservation.[5, 6] In contrast to most other cytoplasmic protein tyrosine kinases, the Janus kinases have no Src homology (SH2 nor SH3) domains.[2] The specific and non-covalent association of these kinases to the intracellular region of cytokine receptors governs their activation upon ligand binding.[3] The JH domains have been shown to be the parts of the Janus kinases that are associated with the cytoplasmic domains of cytokine receptors.[3, 5, 7] The activation of the Janus kinases is mediated by ligand-induced receptor oligomerisation.[8, 9, 10] The Janus kinases are activated by e.g. the type I interferons (IFNα/β and γ), the interleukins (IL2-7, IL-10 and IL-12), growth hormone (GH), prolactin, erythropoietin (Epo), granulocyte-specific colony-stimulating factor (G-CSF), granulocyte-macrophage colony-stimulating factor (GM-CSF), leukaemia inhibitory factor (LIF) and ciliary neurotrophic factor (CNTF).[2, 5, 11]

Activated Janus kinases autophosphorylate,[3] and phosphorylate the cytokine receptors with which they are associated, providing binding sites for the Signal Transducers and Activators of Transcription (STAT) family of transcription factors.[8] The Jaks catalyse phosphorylation of the STAT proteins (seven isoforms, STAT1-4, STAT5A-B and STAT6),[12] that occurs by transfer of the γ phosphate of adenosine triphosphate (ATP) to the hydroxyl group of a tyrosine residue in the STAT protein. After phosphorylation on tyrosine residues, the STAT molecules form homo- or heterodimers,[9] which are translocated into the nucleus. The STAT proteins then bind to DNA, and activate gene transcription.[2] The Jak-STAT signalling cascade has been shown to contribute to growth and survival of e.g. human multiple myeloma cells,[13] acute lymphoblastic leukaemia[14] and a variety of other malignancies.[15, 16] This makes the Janus kinases potential targets for new cancer therapies. One way to interrupt this signalling cascade is to block the binding of ATP to the tyrosine kinases. ATP analogues are generally non-selective, but the development of inhibitors like STI571[17] shows that ATP binding sites can be used as targets for selective drugs.

At the present time none of the Janus kinases have experimentally determined 3-dimensional (3D) structures.[18, 19] In a recent publication, we predicted the 3D structures of the tyrosine kinase domains of Jak2 and Tyk2 by homology modelling, and suggested functional groups for a selective inhibitor of Tyk2 based on Protein Alpha Shape Similarity Analysis (PASSA).[20] PASSA is a new method for mapping protein binding sites, and is especially suited for protein structures predicted by homology modelling. In PASSA, several models for the same protein are used together with structures of other, related proteins to single out unique features of the target protein. Hence, this method is developed especially for design of selective drugs. In PASSA, the binding sites of the protein structures are compared using gaussian property distributions. Discriminant Partial Least Squares (DPLS) regression is used for the data analysis. DPLS regression is PLS regression, where the dependent variables are indicator variables. These results are combined with results from Multiple Copy Simultaneous Search (MCSS),[21] to suggest functional groups of a selective inhibitor. The use of gaussian functions to describe the protein binding sites makes PASSA especially suited for use with homology modelled structures, since the less detailed representation may be more robust than e.g. force field based methods against small structural errors typically present in homology models. Homology modelling in drug design has recently been reviewed.[22]

In this work, we utilised previously suggested functional groups for a selective Tyk2 inhibitor[20] in a pharmacophore search of the database of the National Cancer Institute (NCI). The resulting structures were tested for binding to Tyk2 by computational docking in a homology model of Tyk2.

Structures having the desired functional groups were also generated by *de novo* ligand design. The most promising drug candidates resulting from this analysis were tested for selectivity towards Tyk2 by computational docking in seven protein kinase structures, in addition to the homology model of Tyk2.

Methods

Pharmacophore search

The 3D structure database of the NCI from August 2000 (http://cactus.nci.nih.gov/) (250241 structures) was searched using the pharmacophore search routines in Molecular Operating Environment (MOE).[23] Functional groups for a selective Tyk2 inhibitor have been proposed earlier by our group,[20] based on MCSS. Selected MCSS fragments[20] defined the pharmacophore.

A match on at least six of the pharmacophore query features (MCSS fragments) was required. A query feature is a point in space with a radius-like tolerance on spatial proximity and an associated expression indicating electrostatic properties. To allow some variation from the MCSS fragments, the following proximity tolerances were used for the different query features: Aromatic (benzene rings): 4.0 Å, hydrophobic (CH_3-groups): 2.0 Å, hydrogen donor or hydrogen acceptor: 1.6 Å. This is about twice the actual size of the fragments. The compounds of the database satisfying at least six of the specified functionalities (having atoms with properties overlapping with the pharmacophore features) were first filtered according to distance from the ATP binding site. All compounds having atoms within 10 Å of the docked conformation of ATP[20] were kept for further analysis.

Computational docking analysis

The hits from the pharmacophore search were docked in a previously reported[20] homology model of Tyk2. Two different docking procedures were used: Docking with MOE-Dock,[23, 25] and docking with a new docking method recently developed by our group.[26] MOE-Dock uses the sum of the electrostatic and the dispersive interaction energy between the ligand and the target and the intramolecular energy of the ligand to rank the structures. The molecular mechanics (MM) force field MMFF94[27] was chosen for the docking study as it predicts both intermolecular hydrogen bonding and geometries of small molecules quite well.[28] A smooth non-bonded cut-off of 10-12 Å was used. In our new docking method, a score function based on gaussian property descriptions is used. Both methods use Tabu search[25] for the geometry search.

Docking with MOE-Dock

ATP has previously been docked into the homology model of Tyk2.[20] The hits from the pharmacophore search having atoms within 10 Å of the docked conformation of ATP were docked into the homology model of Tyk2 as a first screening, using ten MOE-Dock runs of 1000 iterations each. A docking box of 125x125x125 grid points with 0.375 Å spacing between each grid point was used. The docking box was centred on the docked conformation of ATP. All structures from this docking analysis having docking energies <5000 kcal/mol (112 structures) were further docked using ten runs of 25000 iterations each. This threshold of 5000 kcal/mol was chosen based on the distribution of the docking energies.

MOE-Dock uses grid-based potential fields[23] to calculate interaction energies between the ligand and the receptor. This grid-based method calculates the potential energy grids only once, at the beginning of the docking procedure. Hence, protein flexibility is not taken into account in these calculations.

The conformation of each drug candidate having the lowest docking energy was also scored using the gaussian-based score function recently developed by our group.[26]

Docking using the gaussian-based docking method

The structures from the pharmacophore search of the NCI database described above were also docked using our newly developed gaussian-based docking method.[26] The largest and most flexible compounds (917 out of 1168 structures) were removed from the compound set prior to the docking. The compounds were chosen according to Kier flexibility index.[29] The threshold value for the Kier flexibility index was chosen by inspection of the structures.

A docking analysis with 100 Tabu runs of 1000 iterations each was carried out, using a docking box with 3 Å padding around the protein structure. MMFF94 with a smooth non-bonded cut-off of 10-12 Å was used. A threshold value of 500 kcal/mol for the ligand Lennard-Jones potential was used in the geometry search.[26] This docking method is independent of hydrogen atoms and partial charges.

De novo ligand design

LigBuilder[30] was used to design new structures having the required functional groups. Structures were built using selected molecular fragments from previous MCSS results[20] as "seed" structures in the "GROW" function of LigBuilder. The binding pocket of Tyk2 was defined by the MCSS fragments defining the Tyk2 pharmacophore. The resulting structures were energy minimised in MOE[23] (100 iterations with steepest descent, 100 iterations with conjugate gradient and 200 iterations with Truncated Newton optimisation) in complex with the homology model of Tyk2 with all receptor atoms fixed. The force field MMFF94[27] with implicit solvation was used. Following the energy minimisation, the structures were ranked according to binding affinities estimated using the gaussian-based score function.[26]

The compounds in the NCI database were searched for similarity to the most promising structures from the *de novo* ligand design, using the pharmacophore search routines in MOE.[23, 24] To approximate the size of the functional groups of the ligands, the following proximity tolerances were used for the different pharmacophore query features: Aromatic rings: 2.0 Å, hydrophobic groups: 1.8 Å, hydrogen donor or hydrogen acceptor: 0.8 Å.

Testing of promising drug candidates for selectivity to Tyk2

To test the most promising drug candidates from the pharmacophore search and the *de novo* ligand design for selectivity towards Tyk2, they were docked in the following kinase structures, in addition to the homology models of Tyk2 and Jak2[20]: RCSB Protein Data Bank (PDB)[18,19] entries 1ir3 (Insulin-receptor tyrosine kinase), 1byg (C-terminal Src kinase), 1fgk (tyrosine kinase domain of Fibroblast growth factor receptor 1), 1fpu (Abelson (Abl) kinase), 1qcf (Haematopoetic cell kinase (Hck)) and 1qpc (Lymphocyte-specific kinase (Lck)). The protein structures were aligned to the homology model of Tyk2 in MOE prior to docking. A modified version of the Needleman and Wunsch approach[31] with a structural correction and the Blosum 62[32] similarity matrix was used for the sequence alignments. The 3D structures were superposed as described by Shapiro *et al.*[33] The docking analysis was performed with our gaussian-based docking method, as described above.

Results and discussion

Pharmacophore search and docking

The Tyk2 pharmacophore used for searching the NCI database (and for *de novo* ligand design) was based on previously selected molecular fragments from MCSS.[20] These fragments are shown together with the docked conformation of ATP[20] in Figure 1.

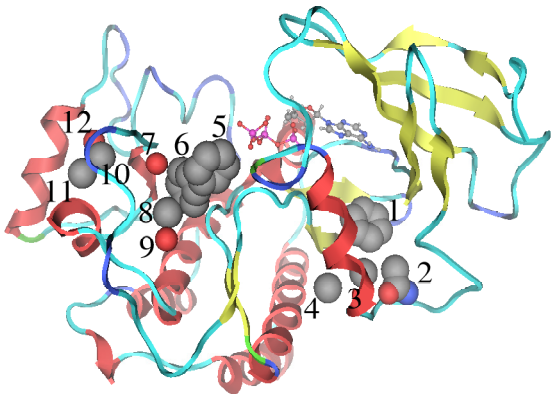


Figure 1. The Tyk2 pharmacophore used for the database search and *de novo* ligand design. The pharmacophore was defined by fragments from MCSS.[20] The docked conformation of ATP[20] is also shown.

Pharmacophore searching of the NCI database resulted in 1168 compounds having properties that matched at least six of the specified functionalities, and were placed within 10 Å of the docked conformation of ATP.

Docking with MOE-Dock

The five compounds from the NCI database that were predicted to have the lowest docking energy by MOE-Dock are listed in Table 1. The docking energies are shown together with the estimated binding affinity to Tyk2 predicted using our gaussian-based score function. There is good correlation between the docking energies from MOE-Dock and the binding affinity estimated using our gaussian-based score function for all compounds in Table 1, except for the compound with NSC number 27773.

Table 1. Results from computational docking of selected structures from the NCI database with MOE-Dock.

NSC number	Docking energy from MOE-Dock (kJ/mol)	Estimated binding affinity[a] to Tyk2 (kJ/mol)
40148	-323.0	-2.31
159203	-113.0	-1.20
3766	-37.1	-0.01
29377	-3.31	-0.03
27773	26.3	-4.59

[a] The binding affinity was predicted using our gaussian-based score function.[26]

Comparison of the placement of these ligands in the Tyk2 binding site with the Tyk2 pharmacophore showed that none of the docked conformations of the ligand structures had functional groups completely overlapping with the pharmacophore. The ligand structure with NSC number 40148 is also small, and very flexible. The docked structure of this ligand had groups overlapping with the oxygen-containing sugar ring in the docked conformation of ATP. It is

therefore not likely to be Tyk2 selective. The hydrophobic rings of ligand structure 159203 were placed close to benzene rings "5" and "6" from Figure 1, but the ring structures were not overlapping. This may, however, be caused by inaccuracies in the docking analysis. This structure is therefore considered to be a possible drug candidate, in spite of the low estimated binding affinity. The same is true for the ligand structures with NSC numbers 3766 and 29377. The hydrophobic part of ligand structure 27773 was overlapping with benzene ring "5", but in the same way as 40148, this ligand is small and flexible, and therefore not likely to be selective to Tyk2.

The mean experimental binding affinity for the set of structures used to train the gaussian-based docking method[26] was –35 kJ/mol. None of the docked structures from MOE-Dock had estimated binding affinities below –35 kJ/mol. This indicates that even though some of the structures in the NCI database have functional groups that match the Tyk2 pharmacophore, they may not bind very strongly to Tyk2.

For comparison, the binding affinity was also estimated in the same way for six different X-ray structures of protein kinases in complex with known ligands from the PDB.[18,19] Some of these X-ray structures were used as templates in the homology modelling of Tyk2.[20] The average estimated binding affinity for these six protein kinase complexes was -18 kJ/mol (Table 2). None of the compounds from the NCI database shown in Table 1 had estimated binding affinities below -18 kJ/mol. However, when the 112 docked structures from the last MOE-Dock screening were sorted according to binding affinities to Tyk2 estimated using our gaussian-based score function (keeping the ligand conformations produced by MOE-Dock), three of the compounds had binding affinities below –18 kJ/mol (Table 3). The docked conformations of 116725 and 167941 were both placed close to benzene rings "5" and "6", while the docked conformation of 231503 was placed close to fragments "1"-"4" from Figure 1.

Table 2. Estimated binding affinities for protein kinases in complex with known ligands (from experimental structures).

PDB entry	Ligand	Estimated binding affinity[a] (kJ/mol)
1agw	SU4984	-15.9
1fpu	PRC	-29.3
1iep	STI571	-35.3
1ir3	ANP-Mg	-13.0
1k3a	ACP	-9.31
1qpc	ANP	-7.90

[a] The binding affinity was predicted using our gaussian-based score function.[26]

Table 3. The three compounds from the NCI database having estimated binding affinities to Tyk2 below –18 kJ/mol.

NSC number	Docking energy from MOE-Dock (kJ/mol)	Estimated binding affinity[a] to Tyk2 (kJ/mol)
116725	49.8	-21.7
167941	512.1	-20.5
231503	3073.1	-18.2

[a] The binding affinity was predicted using our gaussian-based score function.[26]

Docking using the gaussian-based docking method

In the same way as for MOE-Dock, docking with the gaussian-based docking method[26] did not identify any compounds from the NCI database with estimated binding affinities to Tyk2 below –35 kJ/mol. However, one of the structures had estimated binding affinity below the average for the six X-ray structures of protein kinase complexes in Table 2 (-18 kJ/mol). The estimated binding affinity to Tyk2 for this compound is given in Table 4. The docked conformations of this compound from MOE-Dock and docking with the gaussian-based docking method were quite similar (Figure 2).

Table 4. Results from computational docking of selected structures from the NCI database with our gaussian-based docking method.[26]

NSC number	Estimated binding affinity to Tyk2 (kJ/mol)
116725	-26.13

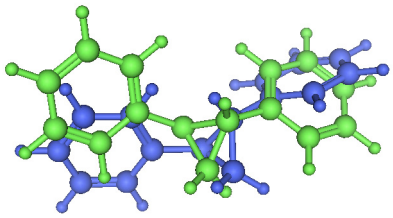


Figure 2. The docked conformations of 116725 produced by MOE-Dock (green) and our gaussian-based docking method (blue).

Since none of the two docking methods used in this work were able to identify any compounds from the NCI database with estimated binding affinities below the average for the set of X-ray structures used to train the gaussian-based docking method, new structures were generated with *de novo* ligand design, in order to find compounds that bind more strongly to Tyk2.

De novo ligand design

In each LigBuilder run, 200 candidate structures were generated. Benzene rings “1”, “5” and “6” (Figure 1) were used separately as “seed” fragments. Two LigBuilder runs with benzene rings “5” and “6”, respectively, and one LigBuilder run with benzene ring “1” were carried out. In total, 1000 structures were generated. Estimation of binding affinities for these structures using the gaussian-based score function,[26] showed that using benzene ring “1” from Figure 1 as “seed” fragment resulted in the most promising drug candidates. In total, 162 of our compounds had predicted binding affinities below the mean experimental binding affinity for the set of structures used to train the gaussian-based docking method[26] (–35 kJ/mol). One of these compounds was generated with benzene ring “5” as “seed” fragment (called “5_1”), while all the other compounds were generated with benzene ring “1” as “seed” fragment (“1_1”-“1_161”). Table 5 shows the estimated binding affinities for the structures generated using benzene ring “1” with estimated affinity below -45 kJ/mol (ten structures), together with the one compound generated with benzene ring “5” as “seed” fragment having estimated affinity below -35 kJ/mol.

Table 5. Estimated binding affinities to Tyk2 for the most promising drug candidates generated by *de novo* ligand design.

Ligand structure	Estimated binding affinity[a] to Tyk2 (kJ/mol)
1_1	-47.60
1_2	-46.94
1_3	-46.89
1_4	-46.88
1_5	-46.15
1_6	-45.73
1_7	-45.52
1_8	-45.47
1_9	-45.44
1_10	-45.44
5_1	-35.99

[a] The binding affinity was predicted using our gaussian-based score function.[26]

The results in Table 5 show that the estimated binding affinities for the structures generated with LigBuilder are much lower than for any of the compounds from the NCI database. Hence, these structures are more likely to be effective as Tyk2 inhibitors. There is, however, no guarantee that they do not bind to other kinases as well. The selectivity of these compounds towards Tyk2 was tested by computational docking.

Testing of promising drug candidates for selectivity to Tyk2

The six most promising structures from the docking analysis with MOE-Dock and our gaussian-based method (159203, 3766, 29377, 116725, 167941 and 231503), together with the eleven structures in Table 5 were docked in seven protein kinase structures, in addition to Tyk2, using the gaussian-based docking method. The estimated binding affinities are shown in Table 6. The gaussian-based docking method was chosen for this study, since it was developed especially for use with homology modelled proteins.[26] The use of gaussian functions gives a less detailed representation that may be more robust than force field based methods against small structural errors typically present in homology models. Homology models of Tyk2 and Jak2 were used here.

Table 6. Estimated binding affinities (kJ/mol) for the most promising drug candidates after docking in seven protein kinase structures in addition to the homology model of Tyk2.

Ligand structure	Estimated binding affinity (kJ/mol)							
	Tyk2	Jak2	1ir3	1byg	1fgk	1fpu	1qcf	1qpc
159203	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3766	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29377	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
116725	-26.13	0.0	-33.06	0.0	0.0	0.0	0.0	0.0
167941	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
231503	-1.73	0.0	-1.31	-1.30	0.0	0.0	0.0	-1.85
"1_1"	-4.35	-6.50	-5.80	-3.91	-6.19	-1.98	-7.98	-4.79
"1_2"	-49.56	0.0	-6.61	0.0	0.0	0.0	-6.86	0.0
"1_3"	-4.46	-6.04	-5.56	-3.81	-5.66	-5.47	-7.93	0.0
"1_4"	-0.013	0.0	0.0	0.0	0.0	0.0	0.0	-3.60
"1_5"	-5.91	-6.66	-5.59	-3.86	-4.63	-5.24	-7.14	-5.06
"1_6"	0.0	0.0	-4.41	0.0	0.0	-1.86	-0.93	0.0
"1_7"	-46.14	-6.27	-5.31	-4.14	-5.87	-5.79	0.0	-5.66
"1_8"	-22.66	0.0	-7.22	0.0	-4.00	0.0	0.0	0.0
"1_9"	-42.56	-5.74	-5.22	-3.76	-4.69	-5.24	-7.08	-4.97
"1_10"	-43.26	-6.89	-5.34	-4.20	-5.0	-0.90	-8.16	-5.47
"5_1"	-4.01	0.0	$-2.0e^{-4}$	0.0	-0.021	0.0	-6.89	0.0

The zero entries in Table 6 are caused by ligand placements outside the grid used to estimate the binding affinity. Hence, these ligands are docked outside the binding pocket of the proteins. Ligands binding outside the active site region are not likely to inhibit activity, and therefore not relevant for this study.

The results in Table 6 indicate that the compound with NSC number 116725 might be a selective inhibitor of Tyk2 and insulin receptor tyrosine kinase. As shown in Figure 3 A, the docked structure of 116725 overlap with benzene rings "5" and "6" from the Tyk2 pharmacophore shown in Figure 1. Hence, this compound may be a promising drug candidate. The docked conformation of 116725 in the homology model of Tyk2 is shown in Figure 3 B. Figure 4 shows the docked conformation of this compound in insulin receptor tyrosine kinase, together with the ligand in PDB entry 1ir3.

A **B**

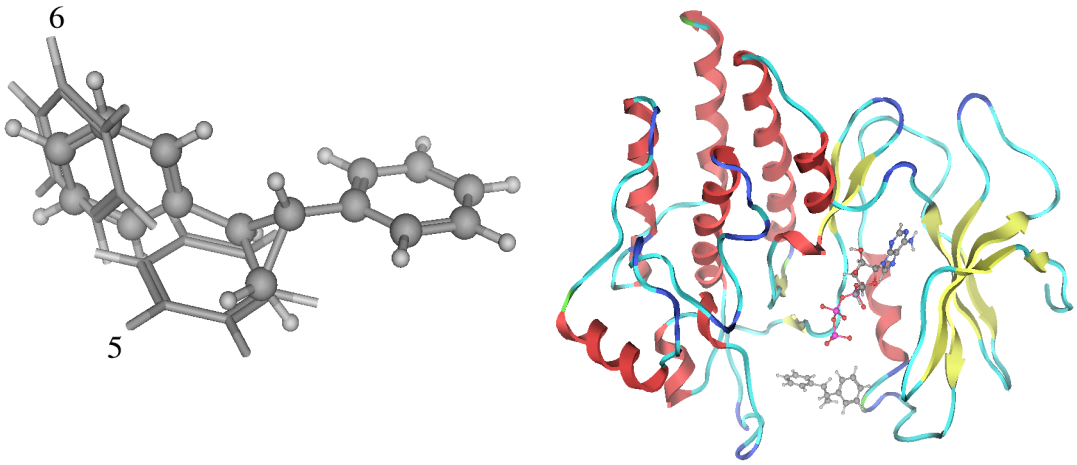


Figure 3. A: The docked conformation of 116725 in the homology model of Tyk2, together with benzene rings "5" and "6" from the Tyk2 pharmacophore. The ligand is rendered in "ball and stick", while the fragments are rendered in "stick".
B: The docked conformation of 116725 in the homology model of Tyk2, together with the docked conformation of ATP.[20]

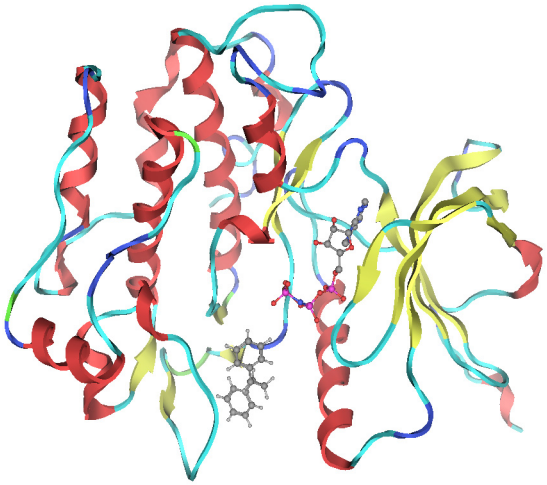


Figure 4. The docked conformation of 116725 in insulin receptor tyrosine kinase, together with the X-ray structure of insulin receptor tyrosine kinase in complex with ANP-Mg (PDB entry 1ir3).

Figure 3 B and Figure 4 show that the orientation of 116725 is different in insulin receptor tyrosine kinase compared to Tyk2, but this compound utilises the same pocket in both structures. Hence, this pocket may not be the best choice in the design of a selective Tyk2 inhibitor. Compounds that bind in the same pocket as fragments "1"-"4" may be more promising. The much lower estimated binding affinity for the structures generated with LigBuilder using benzene ring "1" as "seed" fragment (Table 5), and the fact that compound "5_1" is not Tyk2 selective according to the results in Table 6, support this assumption. Compound 231503 binds in this pocket according to the MOE-Dock study, but the results from docking with the gaussian-based method indicated the contrary. According to the results presented in Table 6, this compound is not Tyk2 selective. None of the other compounds from the NCI database bind in this pocket. Hence, the structures generated with LigBuilder may be more promising as drug candidates.

The primary template used in the homology modelling of Tyk2[20] was the X-ray structure in PDB entry 1qpc. A common problem with homology modelling is that the model is more similar to the primary template than to the target protein.[34] The results in Table 6 indicate that this is not the case for our homology model of Tyk2, since there is no correlation between the estimated binding affinities for 1qpc and Tyk2. If these protein structures were very similar, one would expect the same compounds to bind to both proteins.

The results from our docking analysis indicate that five of the structures generated with LigBuilder are selective inhibitors of Tyk2. Figures 5 - 9 show the docked conformations of these compounds in the homology model of Tyk2, together with fragments from the Tyk2 pharmacophore and the docked conformation of ATP.[20]

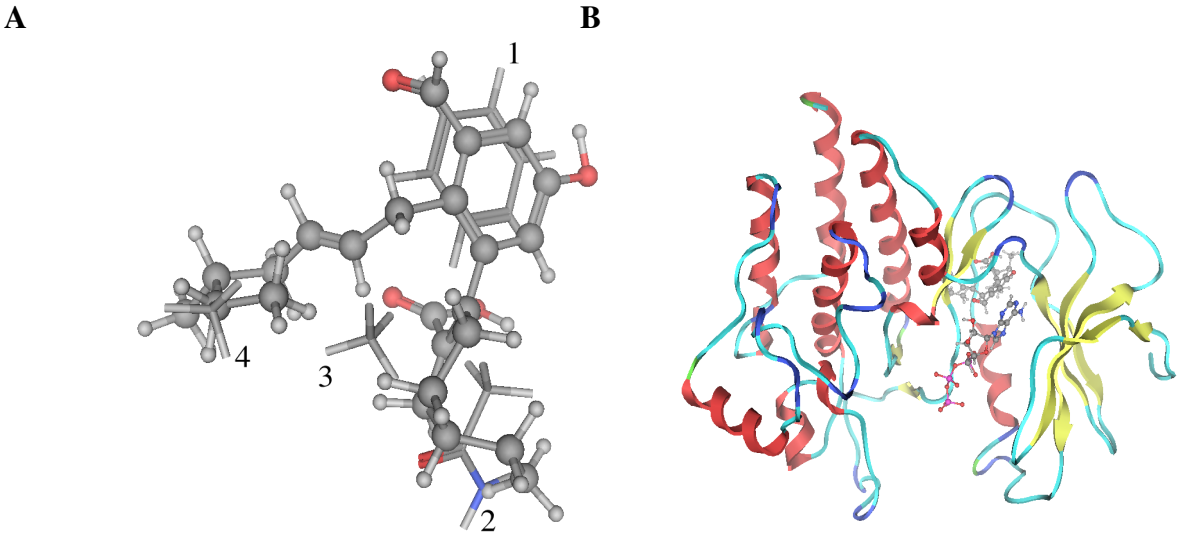


Figure 5. A: The docked conformation of "1_2" in the homology model of Tyk2, together with fragments "1"-"4" from the Tyk2 pharmacophore. The ligand is rendered in "ball and stick", while the fragments are rendered in "stick".
B: The docked conformation of "1_2" in the homology model of Tyk2, together with the docked conformation of ATP.

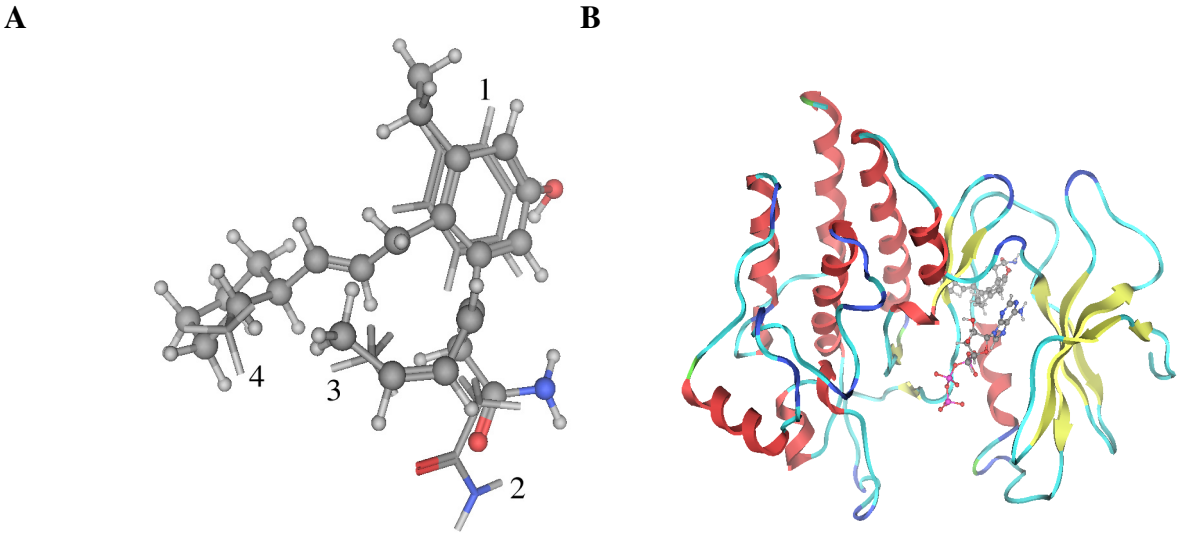


Figure 6. A: The docked conformation of "1_7" in the homology model of Tyk2, together with fragments "1"-"4" from the Tyk2 pharmacophore. The ligand is rendered in "ball and stick", while the fragments are rendered in "stick".
B: The docked conformation of "1_7" in the homology model of Tyk2, together with the docked conformation of ATP.

A **B**

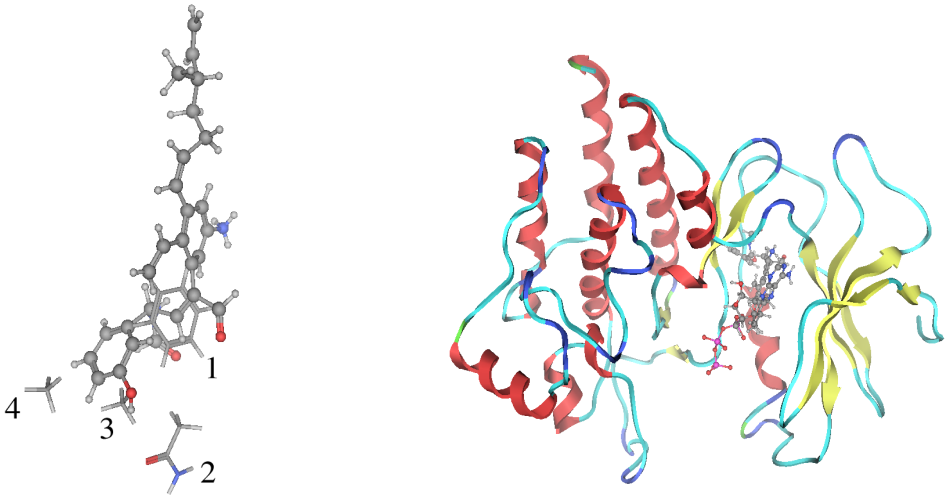


Figure 7. A: The docked conformation of "1_8" in the homology model of Tyk2, together with fragments "1"-"4" from the Tyk2 pharmacophore. The ligand is rendered in "ball and stick", while the fragments are rendered in "stick".
B: The docked conformation of "1_8" in the homology model of Tyk2, together with the docked conformation of ATP.

A **B**

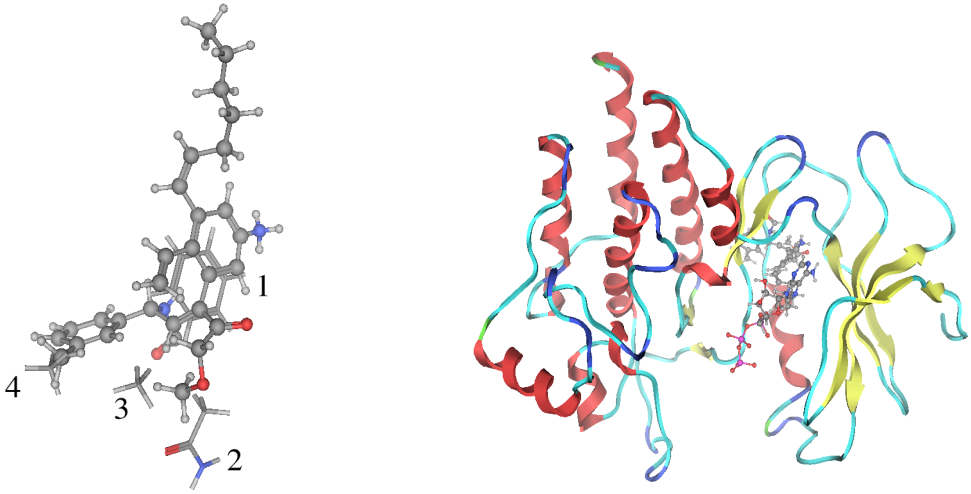


Figure 8. A: The docked conformation of "1_9" in the homology model of Tyk2, together with fragments "1"-"4" from the Tyk2 pharmacophore. The ligand is rendered in "ball and stick", while the fragments are rendered in "stick".
B: The docked conformation of "1_9" in the homology model of Tyk2, together with the docked conformation of ATP.

A **B**

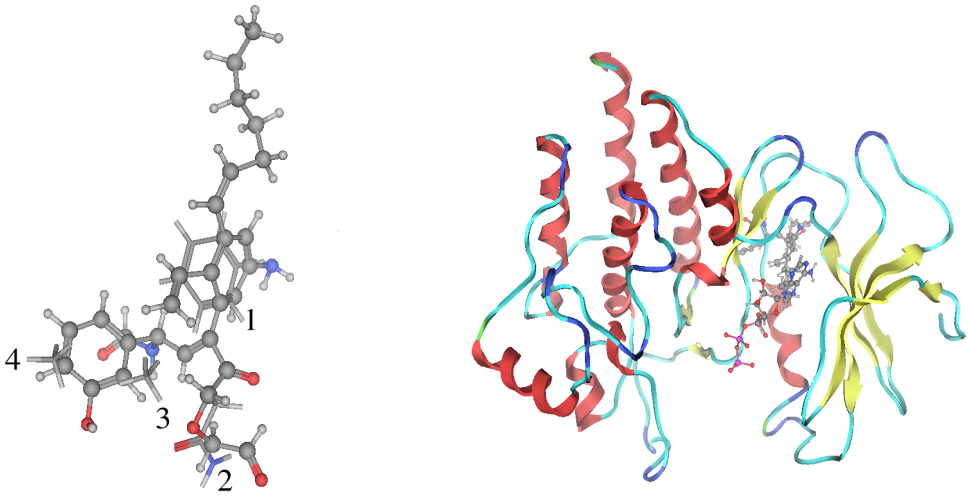


Figure 9. A: The docked conformation of "1_10" in the homology model of Tyk2, together with fragments "1"-"4" from the Tyk2 pharmacophore. The ligand is rendered in "ball and stick", while the fragments are rendered in "stick". **B:** The docked conformation of "1_10" in the homology model of Tyk2, together with the docked conformation of ATP.

The compounds in the NCI database were searched for similarity to the most promising structures from the *de novo* ligand design, "1_2" and "1_7"-"1_10". A match on the pharmacophoric properties of these structures was found for the fourteen structures in Table 7. These compounds were missed in the original pharmacophore search. In the same way as the compounds in Table 6, these compounds were docked in the homology model of Tyk2 and seven other protein kinase structures. The results are given in Table 7.

Table 7. Estimated binding affinities (kJ/mol) from docking of the compounds in the NCI database resembling the most promising structures from *de novo* ligand design.

NSC number	Resembling structure	Estimated binding affinity (kJ/mol)							
		Tyk2	Jak2	1ir3	1byg	1fgk	1fpu	1qcf	1qpc
340033	"1_2"	-2.22	0.0	-3.84	-1.81	0.0	-4.05	-6.45	-3.53
372408	"1_2"	-4.01	-5.31	-3.72	-2.53	-4.16	-4.17	0.0	-3.69
372452	"1_2"	-3.71	-5.28	0.0	-2.34	-1.02	-4.27	-7.21	-3.66
623329	"1_2"	-3.50	-6.61	-6.71	-2.36	-5.71	-5.95	-8.78	-5.39
624404	"1_2"	-3.96	-18.49	-5.95	-9.17	-3.84	-15.77	-5.16	-8.84
627686	"1_2"	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
629605	"1_7"	0.0	-5.69	-5.54	-3.56	-2.72	0.0	-8.94	-4.90
25585	"1_8"	-2.68	0.0	-0.27	0.0	0.0	-2.67	0.0	-2.79
119957	"1_8"	-0.32	0.0	0.0	0.0	0.0	-8.79	0.0	0.0
138557	"1_8"	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
142574	"1_8"	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
157622	"1_8"	0.0	0.0	-0.66	0.0	0.0	0.0	0.0	0.0
203969	"1_8"	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
633715	"1_9"	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

As the results in Table 7 indicate, none of the compounds in the NCI database found to resemble the most promising structures from the *de novo* ligand design bind selectively to Tyk2. They all have relatively low Tyk2 activity. Compound 624404 binds to Jak2 and Abl kinase (PDB entry 1fpu).

The binding to Jak2 might be an artefact of that the X-ray structure in PDB entry 1fpu was the primary template used for the homology modelling of Jak2.[20]

Conclusion

We have screened the NCI database for compounds binding selectively to Tyk2, using two different docking methods. The results from our docking analysis indicated that none of the structures present in the NCI database can be used to inhibit Tyk2 selectively. Even though the two docking methods did not identify the same compounds as the most active ones, they both produced the same conclusion, namely that there are no promising Tyk2 inhibitors in the NCI database. The main purpose of docking methods is to identify the most active compounds. Most docking methods (as these two) are also trained using X-ray structures of protein-ligand complexes. Hence, internal ranking of inactive compounds is bound to fail, and not interesting for drug design purposes. This may be the reason why the two docking methods ranked the compounds in the NCI database differently. However, our analysis provides useful information about parts of the structures that may be used as functional groups of a selective inhibitor of Tyk2, and one compound was found to inhibit Tyk2 and insulin receptor tyrosine kinase selectively. Several promising structures were proposed by *de novo* ligand design. These were tested for selectivity towards Tyk2 by computational docking in seven protein kinase structures, in addition to Tyk2. This study indicated that five of the generated structures might be potential selective inhibitors of Tyk2.

Acknowledgements

We thank Endre Anderssen at the Department of Chemistry, and Anders Sundan and Magne Børset at the Department of Cancer Research and Molecular Biology at The Norwegian University of Science and Technology for helpful discussions. We also thank The Norwegian Research Council for financial support.

References

1. Johnson, L.N.; Noble, M.E.M.; Owen, DJ. Active and Inactive Protein Kinases: Structural Basis for Regulation. *Cell* **1996**, *85*, 149-158.
2. Ihle, J.N.; Witthuhn, B.A.; Quelle, F.W.; Yamamoto, K.; Silvennoinen, O. Signaling through the hematopoietic cytokine receptors. *Annu. Rev. Immunol.* **1995**, *13*, 369-398.
3. Pellegrini, S.; Dusanter-Fourt, I. The structure, regulation and function of the Janus kinases (JAKs) and the signal transducers and activators of transcription (STATs). *Eur. J. Biochem.* **1997**, *48*, 615-633.
4. Van der Geer, P.; Hunter, T.; Lindberg, R.A. Receptor protein-tyrosine kinases and their signal-transduction pathways. *Annu. Rev. Cell Biol.* **1994**, *10*, 251-337.
5. Richter, M.F.; Dumenil, G.; Uze, G.; Fellous, M.; Pellegrini, S. Specific contribution of Tyk2 JH regions to the binding and the expression of the interferon alpha/beta receptor component IFNAR1. *J. Biol. Chem.* **1998**, *273*, 24723-24729.
6. Harpur, A.G.; Andres, A.C.; Ziemiecki, A.; Aston, R.R.; Wilks, A.F. Jak2, a 3rd member of the Jak family of protein tyrosine kinases. *Oncogene* **1992**, *7*, 1347-1353.
7. Yan, H.; Piazza, F.; Krishnan, K.; Pine, R.; Krolewski, JJ. Definition of the Interferon-α Receptor-binding Domain on the TYK2 kinase. *J. Biol. Chem.* **1998**, *273*, 4046-4051.
8. Hubbard, S.R.; Till, J.H. Protein Tyrosine Kinase Structure and Function. *Annu. Rev. Biochem.* **2000**, *69*, 373-398.
9. Heldin, C.H. Dimerization of cell-surface receptors in signal-transduction. *Cell* **1995**, *80*, 213-223.
10. Schlessinger, J.; Ullrich, A. Growth-factor signaling by receptor tyrosine kinases. *Neuron* **1992**, *9*, 383-391.
11. Carter-Su, C.; Smit, L.S. Signaling via JAK tyrosine kinases: Growth hormone receptor as a model system. *Recent. Prog. Horm. Res.* **1998**, *53*, 61-83.
12. Xuan, Y.T.; Guo, Y.R.; Han, H.; Zhu, Y.Q.; Bolli, R. An essential role of the JAK-STAT pathway in ischemic preconditioning. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 9050-9055.
13. Anderson, K. Advances in the Biology of Multiple Myeloma: Therapeutic Applications. *Semin. Oncol.* **1999**, *26*, 10-22.
14. Meydan, N.; Grunberger, T.; Dadi, H.; Shahar, M.; Arpaia, E.; Lapidot, Z.; Leeder, J.S.; Freedman, M.; Cohen, A.; Gazit, A.; Levitzki, A.; Roifman, C.M. Inhibition of acute lymphoblastic leukaemia by a Jak-2 inhibitor. *Nature* **1996**, *379*, 645-648.
15. Wang, L.H.; Kirken, R.A.; Erwin, R.A.; Yu, C.R.; Farrar, W.L. JAK3, STAT, and MAPK signaling pathways as novel molecular targets for the tyrphostin AG-490 regulation of IL-2-mediated T cell response. *J. Immunol.* **1999**, *162*, 3897-3904.
16. Lindauer, K.; Loerting, T.; Liedl, K.R.; Kroemer, R.T. Prediction of the structure of human Janus kinase 2 (JAK2) comprising the two carboxy-terminal domains reveals a mechanism for autoregulation. *Protein Eng.* **2001**, *14*, 27-37.
17. Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, W.T.; Clarkson, B.; Kuriyan, J. Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase. *Science* **2000**, *289*, 1938-1942.
18. The RCSB Protein Data Bank, http://www.rcsb.org/pdb/.
19. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.
20. Tøndel, K.; Anderssen, E.; Drabløs, F. Protein Alpha Shape Similarity Analysis (PASSA): A new method for mapping protein binding sites. Application in the design of a selective inhibitor of Tyrosine kinase 2. *J. Comput. Aided Mol. Des.* **2002**, *16*, 831-840.
21. Miranker, A.; Karplus, M. Functionality maps of binding-sites - a multiple copy simultaneous search method. *Proteins Struct. Func. Genet.* **1991**, *11*, 29-34.
22. Wieman, H.; Tøndel, K.; Anderssen, E.; Drabløs, F. Homology-based modelling of targets for rational drug design. *Mini-Reviews in Medicinal Chemistry* **2004**, In Press.
23. Molecular Operating Environment™, Version 2002.03, *Chemical Computing Group, Inc.*, 2002.

24. Bush, B.L.; Sheridan, R.P. PATTY - a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comp. Sci.* **1993**, *33*, 756-762.
25. Baxter, C.A.; Murray, C.W.; Clark, D.E.; Westhead, D.R.; Eldridge, M.D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins Struct. Funct. Genet.* **1998**, *33*, 367-382.
26. Tøndel, K.; Anderssen, E.; Drabløs, F. A new gaussian-based docking method suitable for use with homology modelled proteins. Unpublished results.
27. Halgren, T.A. Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comp. Chem.* **1996**, *17*, 490-519.
28. Halgren T.A. MMFFVII. Characterisation of MMFF94, MMFF94s and Other Widely Available Force Fields for Conformational Energies and for Intermolecular Interaction Energies and Geometries. *J. Comp. Chem.* **1999**, *20*, 730-748.
29. Hall, L.H.; Kier, L.B. The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D.B., Eds.; VCH Publishers: New York, 1991, 2; pp. 367-422.
30. Wang, R.; Gao, Y.; Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *J. Mol. Model.* **2000**, *6*, 498-516.
31. Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443-453.
32. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915-10919.
33. Shapiro, A.; Botha, J.D.; Pastore, A.; Lesk, A.M. A method for multiple superposition of structures. *Acta Cryst.* **1992**, *A48*, 11-14.
34. Marti-Renom, M.A.; Stuart, A.C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291-325.

Paper VII

Application of Protein Alpha Shape Similarity Analysis (PASSA) in modelling selectivity

Endre Anderssen*, Kristin Tøndel

Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

Abstract

A new method for exploration of protein binding site properties, called Protein Alpha Shape Similarity Analysis (PASSA), has recently been developed. In this work, PASSA has been used to map the properties of the adenosine triphosphate (ATP) binding pockets of proteins in the Protein Kinase C – subfamily and the kinase domains of Abelson (Abl) kinase, Epidermal growth factor receptor (EGFR), Platelet-derived growth factor receptor (PDGFR), c-Src and Protein Kinase A (PKA). Empirical models are developed which are able to predict IC_{50} values for a number of inhibitors towards these proteins, using the data produced by PASSA. Two datasets were analysed by this method and cross-validated correlations (q^2) of 0.75 and 0.62 were obtained. The model parameters can also be examined graphically and give insight into the structural basis for selectivity. Hence, this method is a useful tool for revealing structural features that contribute to selectivity.

Key Words: Partial Least Squares Regression (PLSR), Protein Alpha Shape Similarity Analysis (PASSA), protein kinases, Quantitative Structure-Activity Relationship (QSAR), selectivity.

* Corresponding author. Phone: +47 73 59 41 81. Fax: +47 73 59 16 76. E-mail: endrea@phys.chem.ntnu.no.

Introduction

Drug development is extremely costly and the withdrawal of an otherwise promising drug candidate due to side effects is an enormous waste of resources. Often side effects are caused by interactions with a receptor other than the intended target. Examples of this include aspirin, which targets the protein COX-2, but may cause irritation of the stomach mucosa by inhibition of COX-1 [1]. The protein kinase inhibitor staurosporine, targets the adenosine triphosphate (ATP) binding site of Protein Kinase C. At high concentrations, staurosporine looses its specificity and is toxic [2]. Hence, when promising compounds have been identified, it is of interest to predict their affinity to a number of related proteins as well as the target. This is the inverse of the normal virtual screening problem [3], as the goal is not to screen a large number of drug candidates against a single target, but rather to score a small number of potential drugs against a protein family. Particularly in the case of protein kinase inhibitors targeting the ATP binding site, such information is valuable. ATP binding pockets all have the same overall structure. It is therefore expected that a compound binding to one ATP binding site will bind to a number of other ATP sites in related protein kinases as well. For this reason, use of docking for virtual screening for side effects of protein kinase inhibitors has been suggested [4]. However, reliable docking is computationally expensive, time consuming, and requires protein models of high quality. As virtual screening of a full protein family will most likely require use of homology models to obtain at least some of the protein structures needed, docking may not be an optimal choice of methodology. Though special docking procedures for use with homology models and other low quality protein structures have been developed [5, 6, 7], these methods have not been extensively tested.

A supplement to docking when predicting drug-receptor interactions is the use of three-dimensional (3D) Quantitative Structure-Activity Relationship (QSAR) models such as Comparative Molecular Field Analysis (CoMFA) [8]. QSAR models are empirical models that predict the biological activity of a set of related molecules. Typically, a calibration or training set of a few (30-100) molecules is used. If the modelling is successful, the biological activity of a large number of related molecules may be predicted. The topic of this paper is use of a method related to QSAR modelling in screening for potential off target affinities. An empirical model is made using the recently developed method Protein Alpha Shape Similarity Analysis (PASSA) [9]. With PASSA, the structural properties of a protein binding site can be related to the proteins affinity for a ligand. This is particularly relevant as the number of proteins to screen is relatively small. Even the largest protein families in the human genome typically have less than a thousand members [10]. Previous uses of protein structure information in QSAR work have focused on predicting the affinity towards ‘new’ ligands for a very small number of proteins. Receptor models have been used to guide molecular alignment before 3D QSAR modelling [11], or sequence information about the targets has been included as additional variables in the descriptor data [12]. When screening for side effects, the aim is instead to consider many proteins affinity towards a few ligands. Such data sets are rare in the literature, but can be obtained commercially in the case of kinase inhibitors [13].

Modelling the affinity of a drug for a number of proteins requires a good description of the protein structure and response data to use for calibration. The response data may come from experiments, or from high quality dynamic docking simulations. The standard high throughput docking simulations may not be of sufficient quality to be useful, but free energy perturbation methods have recently been improved to the point where reasonably accurate predictions of binding free energies are possible, though at considerable computational expense [14].

PASSA has been shown to be useful by correctly identifying the reason for the selectivity of STI-571 towards Abelson (Abl) kinase, as well as by aiding the design of a selective Tyk2 inhibitor [9]. The protein binding site representation in PASSA is derived using geometric objects known as alpha spheres. Alpha spheres tend to cluster in ligand binding regions of proteins [15]. A molecular similarity field is computed as a sum of gaussians centred on the alpha spheres and on the protein

atoms. Gaussians centred on alpha spheres are classified as either hydrophobic or hydrophilic, depending on the nature of the protein atoms contacting the alpha sphere. Separate fields are then computed for the density of either hydrophobic or hydrophilic alpha spheres. These fields are referred to as gaussian property fields. The fields are sampled on a set of grid points spanning the binding pocket and the resulting gaussian property fields are used in the data analysis.

In this work we use PASSA to predict biological activity by using the gaussian property field data as independent variables in a Partial Least Squares Regression (PLSR). This regression model may then be able to predict the affinity of the ligands for related proteins based on the alpha sphere densities of their ligand binding sites. Using PASSA to model selectivity within a protein family is a useful supplement to virtual screening with computational docking. Empirical docking methods are trained on diverse sets of compounds and are intended to be as general as possible. Hence, the ability to predict binding modes and binding affinities for a certain protein-ligand complex depends on the similarity of the complex to the structures used to train the docking method. Using PASSA to model selectivity within a protein family, as in this work, allows for more detailed and family-specific modelling of protein-ligand interactions. This method also allows for effective visualisation of the molecular basis for selectivity.

In the work presented here, PASSA has been used to model selectivity of ligands towards two different sets of protein kinases. One is a set of eight Protein Kinase C (PKC) isozymes [16], while the other set consists of structures of the kinase domains of Abl kinase, Epidermal growth factor receptor (EGFR), Platelet-derived growth factor receptor (PDGFR), c-Src, Protein Kinase A (PKA), and two isozymes of PKC [17]. Overactivity of some PKC isozymes has been associated with several disease states, e.g. diabetic complications. The protein kinase inhibitor staurosporine, targets the ATP binding site of PKC. In this work, the activity of staurosporine and nine other 14-membered macrocycles towards the eight different PKC isozymes is modelled using PASSA.

Abl kinase has been shown to play an important role in the development of chronic myelogenous leukaemia (CML) and acute lymphocytic leukaemia (ALL) [17, 18]. Binding of ATP to Abl kinase is crucial for its activity. Hence, the ATP binding site of Abl kinase is an attractive drug target. STI-571 is a selective inhibitor of Abl kinase, c-Kit and PDGFR [18]. During the screening for a selective Abl kinase inhibitor, 37 compounds were tested for binding to Abl kinase, EGFR, PDGFR, c-Src, PKA, PKC-α and PKC-δ [17]. In this work, PASSA has been used to describe the selectivity of the compounds in this dataset.

Materials and methods

Binding affinity data

Dataset 1 consists of a set of IC_{50} values for inhibitors of Protein kinase C (PKC) and was obtained from the work of Jirousek *et al.* [16]. The dataset consists of IC_{50} values for ten kinase inhibitors (Fig. 1) on eight PKC isozymes.

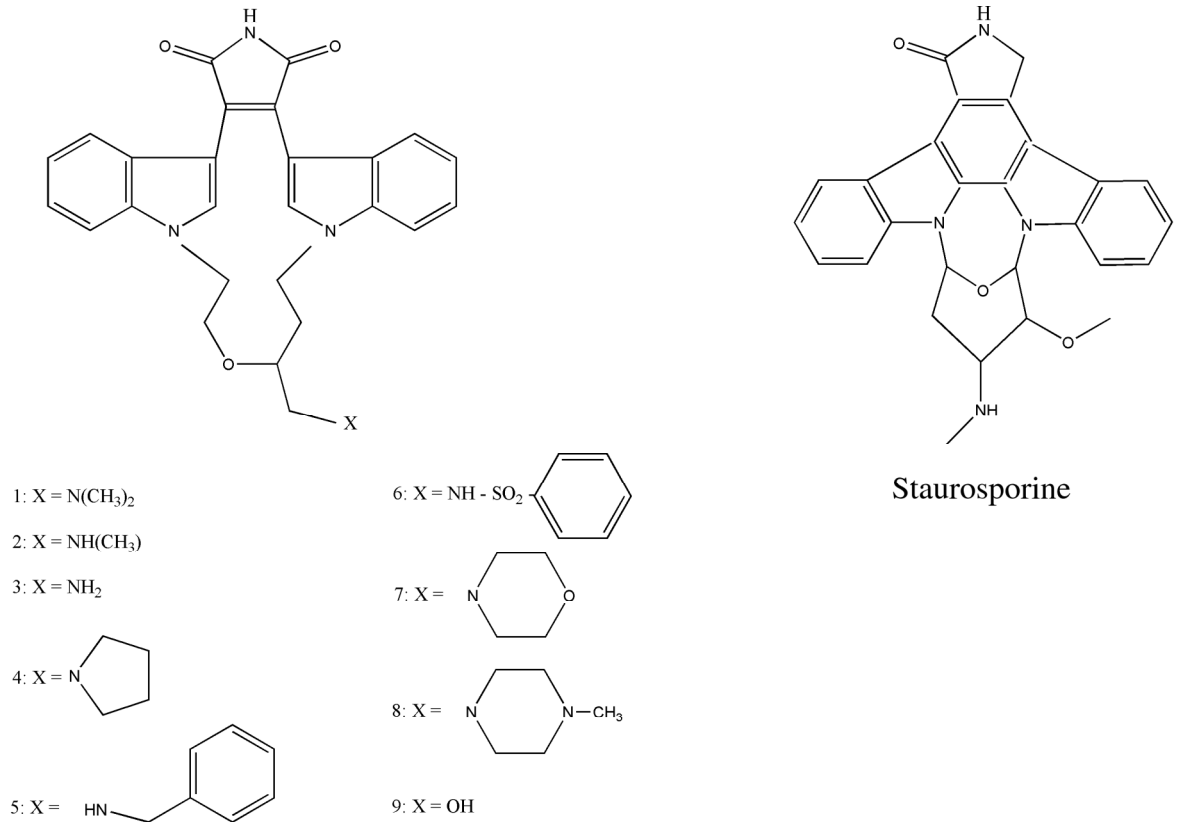


Figure 1. Structures of the ligands used in dataset 1.

Dataset 2 was taken from the work of Zimmermann *et al.* [17]. They tested 37 derivatives of the Abl kinase inhibitor STI-571 for binding to seven different members of the protein kinase family (Abl kinase, EGFR, PDGFR, c-Src, PKA, PKC-α and PKC-δ).

To closer approximate a normal distribution and clarify the interesting variations, both datasets were converted to pIC_{50} values before modelling.

Protein structures

3D structures for the eight PKC isozymes were obtained using homology modelling in Molecular Operating Environment (MOE) [19]. Templates (Table 1) were found with BLAST search in

SwissModel [20]. All templates were aligned using a modified Needleman and Wunsch approach with structural correction and the Blosum 62 similarity matrix [21]. Initial models use the backbone coordinates of the first template. Side chain coordinates are taken from the first conserved residue found amongst all templates. Independent models of the target protein structure were built using a Boltzmann-weighted randomised modelling procedure [22], combined with specialised logic for handling of insertions and deletions [23]. The final model was energy minimised to a root mean square gradient of 0.1 kcal/mol, using the AMBER94 force field [24].

Table 1. Templates used in the homology modelling of the PKC isozymes[α]

Nr	*Target*	*Template 1*	*Template 2*	*Template 3*	*Template 4*	*Template 5*
1	PKC-α	1STC.E	1AO6	1KOB.A	1JKL.A	1PHK
2	PKC-β I	1STC.E	1AO6	1JKL.A	1KOB.A	1FGK.A
3	PKC-β II	1STC.E	1AO6	1JKL.A	1KOB.A	1FGK.A
4	PKC-γ	1STC.E	1AO6	1JKL.A	1PHK	1F3M.C
5	PKC-δ	1STC.E	1AO6	1JKL.A	1F3M	1PHK
6	PKC-ε	1STC.E	1AO6	1F3M.C	1PHK	1QMZ.C
7	PKC-ζ	1PHK	1STC.E	1KOB.A	1AO6	1F3MC
8	PKC-η	1STC.E	1AO6	1PHK	1F3M.C	1QMZ.C

[α] The template structures are named according to RCSB Protein Data Bank (PDB) [25] entries.

Due to a large gap in the sequence alignment between the amino acid sequence of human PDGFR and related proteins, reliable homology modelling of PDGFR was not possible. Instead, a model of the 3D structure of the tyrosine kinase domain of PDGFR was made using the threading software 3D-PSSM [26].

X-ray structures of the tyrosine kinase domains of Abl kinase, EGFR, c-Src and PKA were obtained from the RCSB Protein Data Bank (PDB) [25]. All available PDB entries of these proteins were examined, and those that contained no missing residues in the ATP binding pocket were used in the modelling (Table 2).

Table 2. RCSB Protein Data Bank (PDB) entries used in the modelling.

Nr	*Protein*	*Structures obtained from*
1	Abl	PDB entries: 1FPU, 1IEP, 1M52, 1APM
2	PKA	PDB entries: 1APM, 1CDK, 1FMO, 1JBP, 1JLU, 1L3R, 1Q24, 1YDR, 1YDS, 1YDT, 1CPK
3	EGFR	PDB entries: 1M14, 1M17
4	c-Src	PDB entry: 1BYG
5	PKC-α	Homology modelling (Table 1).
6	PKC-δ	Homology modelling (Table 1).
7	PDGFR	Structure obtained by threading (see main text).

Superpositioning of the protein structures

The protein structure models were aligned using the same methodology as in the homology modelling described above. The structures were superpositioned by rigid body searching, minimising the deviations between Cα and Cβ atoms of amino acids with corresponding positions in the sequence alignment. Only the amino acids with atoms at the surface of the binding cavity were used in the superpositioning. Including the Cβ atoms in the superpositioning makes the result more relevant to ligand binding, as information on side chain orientations is used in the superpositioning.

Generating gaussian property fields

Gaussian property fields were computed for each protein on a fixed grid surrounding the ATP binding pockets of the proteins. A spatial resolution of 0.75 Å was used and the grid dimensions were 40×40×50 grid points. The alpha sphere centres are first identified and assigned either a hydrophobic or hydrophilic weight (ω) of +1 on the basis of the atoms touching the alpha spheres. Dummy atoms are placed at the alpha sphere centres. All protein atoms are assigned negative weights (ω) of - 1 for both fields. Both a hydrophilic and a hydrophobic property field are computed using Equation 1.

$$F(q,j) = \sum_{i=1}^{n} \frac{\omega_{ik}}{(\sigma_i \sqrt{2\pi})^3} \cdot e^{\frac{-r_{iq}^2}{2\sigma_i^2}} \qquad (1)$$

F is the value of the gaussian property field in grid point q of molecule j, ω_{ik} is the value of the physicochemical property k of atom i, r_{iq} is the distance between grid point q and atom i and σ_i corresponds to the atomic radius of atom i.

Contributions from both the real protein atoms and the alpha sphere centre dummy atoms are included in the summation. The data was assembled into a molecular similarity matrix with one row for each protein and two columns (hydrophobic and hydrophilic) for each spatial grid point. Each column of this matrix is used as an independent variable in the regression analysis.

Regression analysis

Irrelevant variations are removed from the gaussian property field data by removing variables with a very low standard deviation (std($\mathbf{F_{qk}}$) < 0.1). Grid points with no positive values of F_{qk} for any of the proteins were also removed as these represent points buried in the interior of all proteins. The filtered property field data were used as independent variables in a PLS regression. Due to the small number of proteins used in this analysis, no variable selection or other form of model optimisation was carried out. To simplify the interpretation of the modelling result, the data from all ligands were analysed in the same PLSR2 model.

Results and discussion

Example 1

Gaussian property fields for the eight PKC isozymes were regressed onto the response pIC_{50} data from ten ligands using PLSR. The regression model was validated by full leave one out cross-validation (Fig. 2), and three principal components (PCs) was found to be optimal, resulting in a cross-validated correlation $q^2 = 0.75$. This is comparable to what may be achieved by other methods used to predict binding affinity such as CoMFA or docking [27, 28]. As no model optimisation other than the choice of the number of PLS components is done on the basis of this cross-validation result, the q^2 is a fairly unbiased estimate of the predictive ability of the model [29].

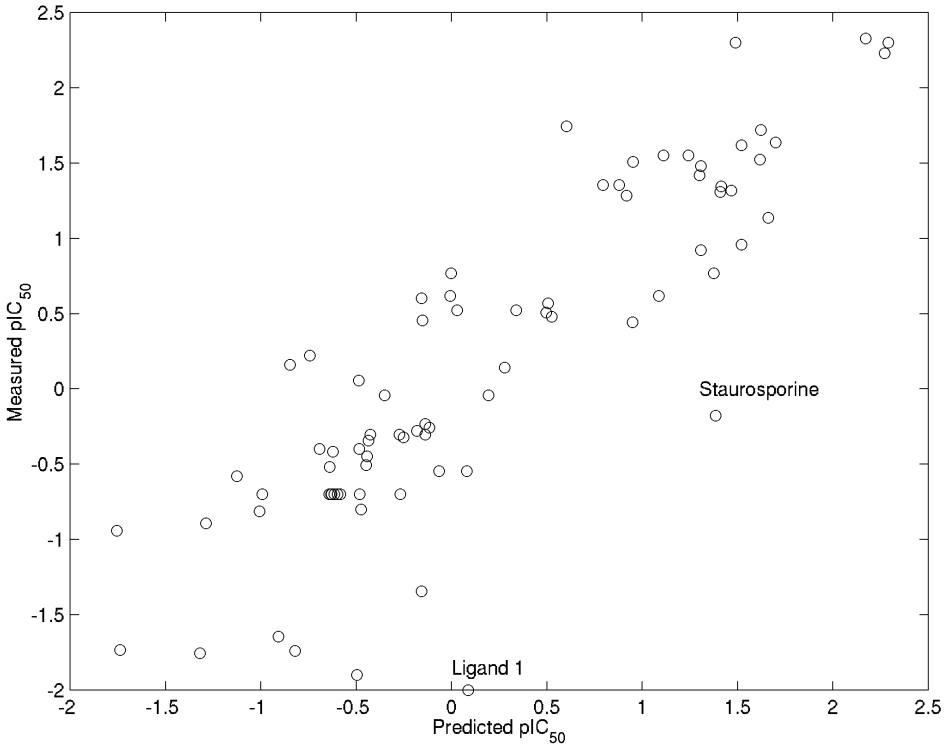


Figure 2. Predicted pIC_{50} from the cross-validation vs. measured pIC_{50} for dataset 1. The marked outliers are the data points of staurosporine and Ligand 1 (Fig. 1) binding to PKC-ζ.

The predicted vs. measured plot (Fig. 2) shows poor predictions for two data points. Both predictions are for the protein PKC-ζ. The homology model for this protein was made using the PDB file 1PHK as primary template, while all the other homology models were made using 1STC. This may cause this homology model to differ from the others purely due to the choice of template, which may influence the predictions for this protein.

Interpretation

The PLSR model produces regression coefficients for each grid point. However, the regression coefficients are representative of one response variable only and may therefore be impractical for a PLSR model with many response variables. A better alternative is to use the loading weights. The loading weights determine the subspace of X used in the PLS regression. Every loading weight vector has an element for every grid point used in the model, and there is one loading weight vector per principal component. These can be visualised in precisely the same way as regression

coefficients. In practice however, the phenomena described by the loading weights themselves are somewhat arbitrary and may not have a physical meaning. Also, as no variable selection is carried out, highlighting relevant loading weights from the model is difficult. An aid in this sense is the scores and loading weights bi-plot. As a three-component model is used, both the scores and the loading weights can be plotted in the same space (Fig. 3). By highlighting interesting regions in the loading weights, the corresponding interaction sites in the proteins can be identified.

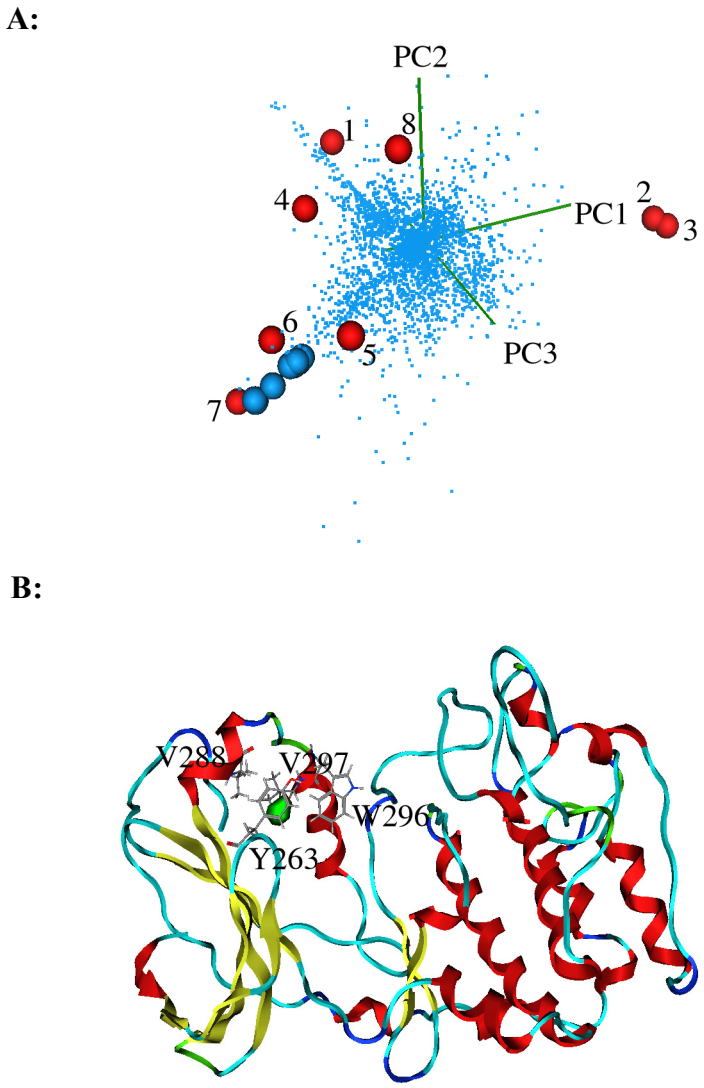


Figure 3. **A:** Scores and loading weight bi-plot. The PLS scores of each protein are shown as red spheres and the PLS loading weights of the grid variables are shown as blue dots. The selected loading weights are rendered as blue spheres. **B:** The structural origin of the selected loading weights. Hydrophobic points are shown in green.

In our case, there are two particularly interesting sets of loading weights. One is protruding towards proteins '1' and '4' and another towards protein '7' (Fig. 3A). Looking at the protrusion towards protein '7' (PKC-ζ), we find that the loading weights spanning this direction correspond to a well-defined hydrophobic area (Fig. 3B).

Example 2

As in example 1, property fields for the proteins were regressed onto the pIC_{50} values for the 37 derivatives of STI-571 (reported in Zimmermann *et al.* [17]) using PLSR. The regression model was validated by full leave one protein out cross-validation, and one principal component was found to be optimal. This resulted in a correlation coefficient (q^2) of 0.62 between the predicted and the measured pIC_{50} (Fig. 4). In this cross-validation the affinity for some ligands towards PDGFR and PKC-α is severely underestimated. In the case of PDGFR, the structure of this protein had to be obtained by threading. The structure may therefore be more uncertain than the structures obtained by experimental X-ray crystallography or homology modelling. The PKC-α structure however, was used successfully in example 1.

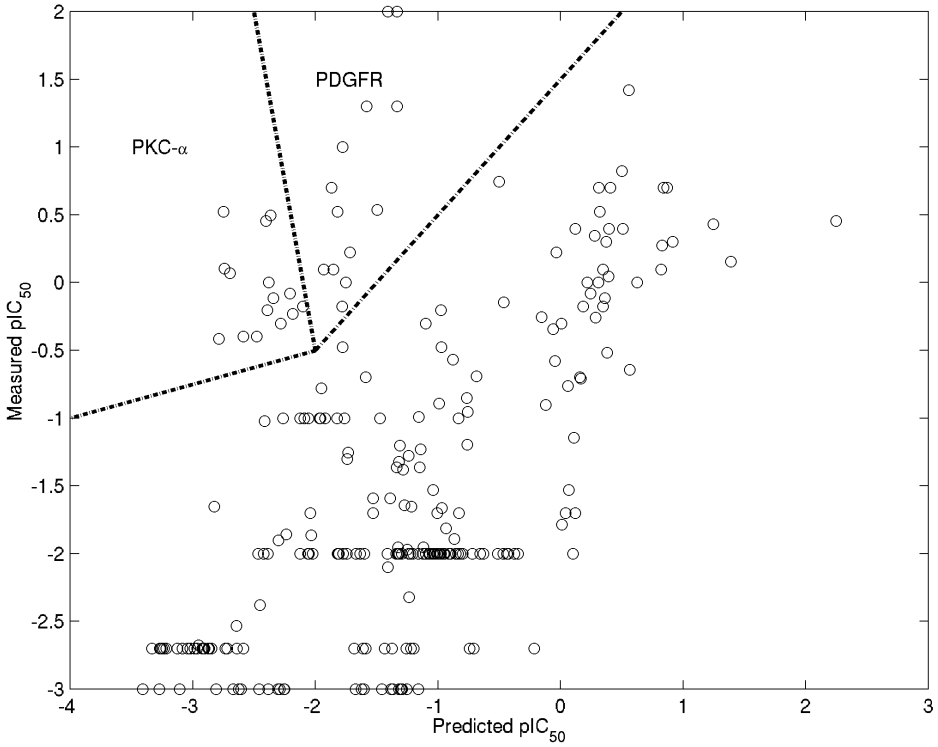


Figure 4. Predicted vs. measured plot from the cross-validation of the model for dataset 2. Regions with outlying samples for PDGFR and PKC-α have been outlined with dashed lines.

In the same way as the loading weights, the regression coefficients from the PLS regression can be mapped back onto the grid points, and structural regions that contribute to selectivity can be identified. As we have experimental structures available for several of the protein-ligand complexes in dataset 2, we are able to test if the results produced by our method resemble the properties of the actual ligands. As an example, the regression coefficients for STI-571 were plotted together with the X-ray structure of Abl kinase in complex with STI-571 (present in PDB [25] entry 1IEP) (Fig. 5). Since STI-571 is known to be a selective inhibitor of Abl kinase, this can be used as a test on how well the results from PASSA correspond to the properties of known, selective inhibitors.

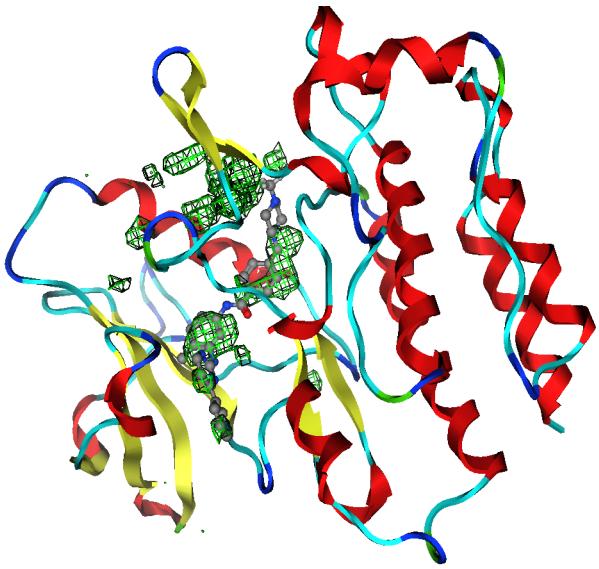


Figure 5. The regression coefficients for the hydrophobicity (green) for STI-571 plotted together with the X-ray structure of Abl kinase in complex with STI-571 (PDB entry 1IEP).

The results in Figure 5 show that the regression coefficients for the hydrophobicity for STI-571 correspond to a large degree with the hydrophobic groups of STI-571. This indicates that PASSA is a useful method for identification of regions in a protein binding site that can be utilised to achieve selective binding of ligands to the protein. However, there are also some regions of high regression coefficients for hydrophobicity that do not overlap with hydrophobic groups on STI-571. This is probably caused by the fact that closely related proteins share many structural features. With the low number of proteins used in this study, correlations between spatial regions that are not directly involved in ligand binding may cause non-binding regions to be highlighted along with binding regions in the analysis.

The examples shown here demonstrate that the PASSA method may be used quantitatively to predict IC_{50} values for a number of ligands within a set of closely related protein targets. The models obtained may also provide insight into regions that are involved in ligand binding to the various proteins. However, due to the small number of proteins used in the modelling, the estimate is still uncertain and more data is needed before any firmer conclusions can be made about the usefulness of this method. If the level of predictive ability obtained in these examples can be expected generally with this method, then the use of PASSA in screening for side effects will be useful.

Acknowledgement

We thank The Norwegian Research Council for financial support.

References

1. Kawai, S., Inflamm. Res., 47 (1998) 102.
2. Sridhar, R., Hanson-Painton, O. and Cooper, D. R., Pharm. Res., 17 (2000) 1345.
3. Shochet, B. K., McGovern, S. L., Binqing, W. and Irwin, J. J., Curr. Opin. Chem. Biol., 6 (2002) 439.
4. Rockey, W. M. and Elcock, A. H., Prot. Struct. Func. Gen., 48 (2002) 664.
5. Schafferhans, A. and Klebe, G., J. Mol. Biol., 307 (2001) 407.
6. McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A. and Brown, F. K., Biopolymers, 68 (2003) 76.
7. Tøndel, K., Anderssen, E. and Drabløs, F., J. Comput. Aided Mol. Des., Submitted.
8. Cramer, R. D., Patterson, D. E. and Bunce, J. D., J. Am. Chem. Soc. 110 (1988) 5959.
9. Tøndel, K., Anderssen, E. and Drabløs, F., J. Comput. Aided Mol. Des. 16 (2002) 831.
10. McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner-McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel, G., Whittaker, A., Holden, J. L., McCann, O. T., Dunham, A., Soderlund, C., Scott, C. E., Bentley, D. R., Schuler, G., Chen, H.-C., Jang, W. H., Green, E. D., Idol, J. R., Maduro, V. V. B., Montgomery, K. T., Lee, E., Miller, A., Emerling, S., Kucherlapati, R., Gibbs, R., Scherer, S., Gorrell, J. H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P. J., Catanese, J. J., Nowak, N., Osoegawa, K., Qin, S. Z., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V. G., Kirsch, I. R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J.-F., Hawkins, T., Myers, R. M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N. E., Cox, D. R., Haussler, D., Kent, W. J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G. P., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X.-N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H.-S., Sakaki, Y., Shimizu, N., Asakawa, S., Kawasaki, K., Sasaki, T., Shintani, A., Shimizu, A., Shibuya, K., Kudoh, J., Minoshima, S., Ramser, J., Seranski, P., Hoff, C., Poustka, A., Reinhardt, R. and Lehrach, H., Nature 409 (2001) 934.
11. Sippl, W., Biorgan. Med. Chem. 10 (2002) 3741.
12. Lapinsh, M., Prusis, P., Lundstedt, T. and Wikberg, J. E. S., Mol. Pharmacol. 61 (2002) 1465.
13. Upstate Group, Inc., 706 Forest Street, Suite 1, Charlottesville, VA 22903.
14. Pearlman, D. A. and Charifson, P. S., J. Med. Chem. 44 (2001) 502.
15. Liang, J., Edelsbrunner, H. and Woodward, C., Protein. Sci. 7 (1998) 1884.
16. Jirousek, M. R., Gillig, J. R., Gonzalez, C. M., Heath, W. F., McDonald, J. H., Neel, D. A., Rito, C. J., Singh, U., Stramm, L. E., Melikian-Badalian, A., Baevsky, M., Ballas, L. M., Hall, S. E., Winneroski, L. L. and Faul, M. M., J. Med. Chem. 39 (1996) 2664.
17. Zimmermann, J., Buchdunger, E., Mett, H., Meyer, T. and Lydon, N. B., Bioorg. Med. Chem. Lett. 7 (1997) 187.
18. Capdeville, R., Buchdunger, E., Zimmermann, J. and Matter, A., Nat. Rev. Drug Disc. 1 (2002) 493.
19. MOE Molecular Operating Environment™, Version 2002.03. Chemical Computing Group Inc., (2002), 1010 Sherbrooke Street West, Suite 910 Montreal, Quebec, Canada H3A 2R7.
20. Peitsch, M.C., Biochem. Soc. Trans. 24 (1996) 274.
21. Heinkoff, S. and Heinkoff, J. G., Proc. Natal. Acad. Sci. 89 (1992) 10915.
22. Levitt, M., J. Mol. Biol. 226 (1992) 507.
23. Fechteler, T., Dengler, U., Schomberg, D., J. Mol. Biol. 253 (1995) 114.
24. Weiner, S. J., Kollman, P. A., Chase, D. A., Singh, U. C, Ghio, C., Alagona, G., Profeta, S., Weiner, P., J. Am. Chem. Soc. 106 (1984) 765.

25. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. Nucleic Acids Res. 28 (2000) 235.
26. Kelley, L. A., MacCallum, R. M. and Sternberg, M. J. E., J. Mol. Biol. 299 (2000) 501.
27. Nicolotti, O., Altomare, C., Pellegrini-Calace, M. and Carotti, A., Curr Top. Med. Chem. 4 (2004) 335.
28. Baxter, C.A., Murray, C.W., Clark, D.E., Westhead, D.R. and Eldridge, M.D., Prot. Struct. Funct. Gen. 33 (1998) 367.
29. Martens, H. A. and Dardenne, P., Chemometr. Intell. Lab. 44 (1998) 99.