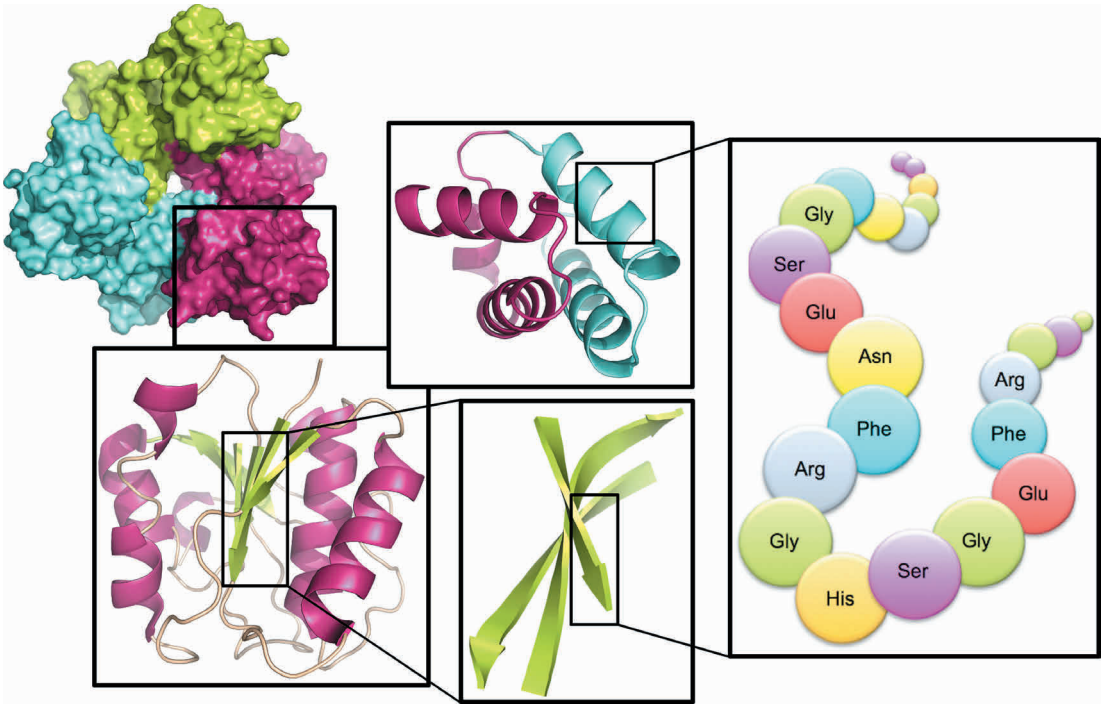
Åbo Akademi University

Käthe M. Dahlström

From Protein Structure to Function with Bioinformatics



Käthe Dahlström

was born on November 10, 1987 in Hanko, Finland. She graduated from Åbo Akademi University in 2011 with a Master of Science in Bioscience. This PhD thesis project in Biochemistry has taken place during 2011 – 2015 under the supervision of Docent Tiina Salminen at the Faculty of Science and Engineering.

From Protein Structure to Function with Bioinformatics

Käthe M. Dahlström

Biochemistry
Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland
2015

From the Faculty of Science and Engineering, Åbo Akademi Univeristy, Åbo Akademi Graduate School & National Doctoral Programme in Informational and Structural Biology

Supervised by

Docent Tiina A. Salminen

Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland

Reviewed by

Professor Antti Poso

Faculty of Health Sciences
University of Eastern Finland
Kuopio, Finland

Dr. Heidi Kidron

Faculty of Pharmacy
University of Helsinki
Helsinki, Finland

Opponent

Professor Bjørn Olav Brandsdal

Department of Chemistry
University of Tromsø
Tromsø, Norway

ISBN 978-952-12-3299-2
Painosalama Oy – Turku, Finland 2015

Ancora Imparo –
I am still learning

Abstract

It has long been known that amino acids are the building blocks for proteins and govern their folding into specific three-dimensional structures. However, the details of this process are still unknown and represent one of the main problems in structural bioinformatics, which is a highly active research area with the focus on the prediction of three-dimensional structure and its relationship to protein function. The protein structure prediction procedure encompasses several different steps from searches and analyses of sequences and structures, through sequence alignment to the creation of the structural model. Careful evaluation and analysis ultimately results in a hypothetical structure, which can be used to study biological phenomena in, for example, research at the molecular level, biotechnology and especially in drug discovery and development.

In this thesis, the structures of five proteins were modeled with template-based methods, which use proteins with known structures (templates) to model related or structurally similar proteins. The resulting models were an important asset for the interpretation and explanation of biological phenomena, such as amino acids and interaction networks that are essential for the function and/or ligand specificity of the studied proteins. The five proteins represent different case studies with their own challenges like varying template availability, which resulted in a different structure prediction process. This thesis presents the techniques and considerations, which should be taken into account in the modeling procedure to overcome limitations and produce a hypothetical and reliable three-dimensional structure. As each project shows, the reliability is highly dependent on the extensive incorporation of experimental data or known literature and, although experimental verification of *in silico* results is always desirable to increase the reliability, the presented projects show that also the experimental studies can greatly benefit from structural models. With the help of *in silico* studies, the experiments can be targeted and precisely designed, thereby saving both money and time. As the programs used in structural bioinformatics are constantly improved and the range of templates increases through structural genomics efforts, the mutual benefits between *in silico* and experimental studies become even more prominent. Hence, reliable models for protein three-dimensional structures achieved through careful planning and thoughtful executions are, and will continue to be, valuable and indispensable sources for structural information to be combined with functional data.

Sammanfattning

Det är sedan länge känt att aminosyror fungerar som byggstenar för proteiner och bestämmer deras specifika tre-dimensionella veckning. Detaljerna kring denna process är dock ännu okända och representerar ett av de största problemen för forskning inom strukturbioinformatik. Denna typ av forskning fokuserar på förutsägning av proteiners tre-dimensionella struktur och dess förhållande till funktionen. Förutsägning av proteinstrukturer omfattar flera faser från sökning och analys av sekvenser och strukturer, till sekvensjämförelse och skapandet av en modellstruktur. Noggrann evaluering och analys av denna resulterar slutligen i en hypotetisk struktur som kan användas för att studera biologiska fenomen inom bl.a. forskning på molekyär nivå, bioteknologi och framför allt inom läkemedelsutveckling.

I denna avhandling modellerades strukturen hos fem olika proteiner med hjälp av templatbaserade metoder, som använder proteiner med känd struktur (templat) för att modellera besläktade eller strukturmässigt liknande proteiner. De resulterande modellerna var en viktig tillgång för tolkningar och förklaringar av biologiska fenomen, såsom vilka aminosyror och samverkningar som är viktiga för funktionen och/eller ligandspecificiteten hos de studerade proteinerna. De fem proteinerna representerar olika fallstudier av växlande svårighetsgrad p.g.a. varierande templattillgång, vilket resulterade i olika processer för strukturmodelleringen. Denna avhandling presenterar tekniker och överväganden som bör beaktas i modelleringsprocessen för att kunna producera en tillförlitlig modellstruktur. Varje projekt visar att tillförlitligheten till hög grad är beroende av omfattande inkorporering av experimentell data eller känd litteratur. Utöver detta är experimentell verifiering av in silico resultat alltid önskvärt för att öka tillförlitligheten, men de presenterade projekten påvisar att även de experimentella studierna kan dra fördel av strukturmodeller. Med hjälp av in silico studier kan experimenten riktas mot ett specifikt mål och planeras i detalj, vilket sparar både pengar och tid vid utförandet av de experimentella studierna. De ömsesidiga fördelarna mellan in silico och experimentella studier blir allt mer framträdande tack vare ständig utveckling och förbättring av programmen och mjukvaran som används inom strukurbioinformatik, samt p.g.a. att templatomfånget utökas. Tillförlitliga modellstrukturer som uppnåtts genom noggrann planering och eftertänksamma utföranden är därmed värdefulla källor för strukturinformation som sedan kan kombineras med funktionell data.

Table of contents

List of original publications

This thesis is based on the following original publications, which are referred to by Roman numerals (I-VI) in the text:

I. Blikstad C., **Dahlström K.M.**, Salminen T.A., Widersten M. (2013) – Stereoselective oxidation of aryl-substituted vicinal diols into chiral α-hydroxy aldehydes by re-engineered propanediol oxidoreductase. *ACS Catalysis* **3**: 3016–3025.

II. Blikstad C., **Dahlström K.M.**, Salminen T.A., Widersten M. (2014) – Substrate scope and selectivity in offspring to an enzyme subjected to directed evolution. *FEBS Journal* **281**(10): 2387-2398.

III. Reinés M., Llobet E., **Dahlström K.M.**, Pérez C., Llompart C., Salminen T.A., Bengoechea J.A. (2012) – Deciphering the acylation pattern of *Yersinia enterocolitica* lipid A. *PLOS Pathogens* **8**(10): e1002978. doi: 10.1371/journal.ppat.1002978.

IV. Llobet E., Martínez-Moliner V., Moranta D., **Dahlström K.M.**, Regueiro V., Tomás A., Cano V., Pérez-Gutiérrez C., Frank C.G., Fernández-Carrasco H., Insua J.L., Salminen T.A., Garmendia J., Bengoechea J.A. (2015) – Deciphering tissue-induced *Klebsiella pneumoniae* lipid A. Accepted to *Proceedings of the National Academy of Sciences of the United States of America*.

V. *Carmel D., ***Dahlström K.M.**, Holmström M., Allahverdiyeva Y., Battchikova N., Aro E.M., Salminen T.A., Mulo P. (2013) – Structural model, physiology and regulation of Slr0006 in *Synechocystis* PCC 6803. *Archives of Microbiology* **195**(10-11): 727-736.

VI. **Dahlström K.M.**, Salminen T.A. (2015) – 3D model for cancerous inhibitor of protein phosphatase 2A armadillo domain unveils highly conserved protein-protein interaction characteristics. *Journal of Theoretical Biology* **386**: 78-88.

*Equal contribution to the work.

Publications are reproduced with the permission of the publishers.

Contributions of the author

The author of this thesis performed all computational work, *i.e.* database searches, sequence alignments, three-dimensional structural modeling, structural analysis and docking studies. The author wrote all sections regarding her own work in publications I-VI.

Additional publications

Lehtimäki N., Koskela M.M., **Dahlström K.M.,** Pakula E., Lintala M., Scholz M., Hippler M., Hanke G.T., Rokka A., Battchikova N., Salminen T.A., Mulo P. (2014) – Post-translational modifications of ferredoxin-NADP+ oxidoreductase in *Arabidopsis thaliana* chloroplasts. *Plant Physiology* **166**(4): 1764-1776.

Edstam M.M., Laurila M., Höglund A., Raman A., **Dahlström K.M.**, Salminen T.A., Edqvist J., Blomqvist K. (2013) – Characterization of the GPI-anchored lipid transfer proteins in the moss *Physcomitrella patens*. *Plant Physiology and Biochemistry* **75C**: 55-69.

Toivola J., Nikkanen L., **Dahlström K.M.**, Salminen T.A., Lepistö A., Vignols H.F., Rintamäki E. (2013) – Overexpression of chloroplast NADPH-dependent thioredoxin reductase in *Arabidopsis* enhances leaf growth and elucidates *in vivo* function of reductase and thioredoxin domains. *Frontiers in Plant Science* **4**(389): doi: 10.3389/fpls.2013.00389.

Acknowledgements

The work for this thesis was carried out at the Structural Bioinformatics Laboratory (SBL), Faculty of Science and Engineering, Åbo Akademi University during 2011 – 2015. Many people have taken part in this journey and helped me along the way. Without you this would not have been possible and as memorable.

First of all I would like to thank my supervisor, Docent *Tiina Salminen*, and Professor *Mark Johnson* for giving me the opportunity to do my thesis in their group. *Tiina*, thank you for patiently sharing your knowledge with me, believing in me and always encouraging me when I need it. Your support and understanding means a lot to me and helps me move forward with my work. I would also like to thank Professor *Antti Poso* and Dr. *Heidi Kidron* for taking the time to read my thesis and for the valuable comments on how to improve it. This work would not have been possible without interesting and fruitful collaborations with other research groups. Therefore, I want to express my gratitude to all my co-authors from the research groups of Professor *José A. Bengoechea* and Assistant Professor *Paulo Mulo*, along with Dr. *Cecilia Blikstad* and Professor *Mikael Widersten*. I would also like to acknowledge the director Professor *Mark Johnson* and coordinator *Fredrik Karlsson* from the National Doctoral Programme in Informational and Structural Biology (ISB). It has been a pleasure belonging to such an excellent graduate school and taking part in the fun and inspiring meetings. I also want to thank my thesis committee members Dr. *Heidi Kidron* and Dr. *Juha Okkeri* for all their valuable ideas and discussions regarding this thesis project.

I would like to thank the past and the present members of SBL for all the help I have received, for the company during lunch and coffee breaks and for the good moments we have shared. I am glad to have had the opportunity to meet you all and I wish you all the best in life. Thank you especially *Fredrik Karlsson* for all the help with practical matters, *Outi Salo-Ahen* for always taking the time for a little chat, *Tomi Airenne* (although you killed my baby cactus) for every day asking me how I am and thanks also to *Jukka Lehtonen* for all the help with computer problems during the years and for patiently trying to teach me about computers and programming despite my very basic knowledge and silly questions. I would also like to thank *Leonor Carvalho* for taking the picture on the inside of the cover of this thesis, the discussions and for the all the fun times we have had traveling. A certain romantic hotel will forever be in my memories and make me laugh. A special thanks to *Eva Bligt-Lindén* for being the best colleague and friend that anyone could ever ask for. This journey has been so much more fun because of you. Our trips to far away countries, our discussions about everything and nothing and all the laughs we have shared are just some of the things I will cherish in my

memories. Most of all I value your real, true and unconditional friendship. Thank you also for allowing me to spend so much time playing with *Edvin*. I bet you have sometimes thought you have two children instead of one.

I would also like to thank the Biochemistry staff for creating a nice work atmosphere and for all the help I have received. I am grateful for the administrative and technical help from *Pirkko Luoma, Eve Hed-Kattelus, Elsmarie Nyman, Jussi Meriluoto* and *Juha-Pekka Sunila*. Thanks also to my other friends at the former Department of Biosciences, especially *Daniela Karlsson, Josefin Halin, Marika Sjöqvist* and *Heidi Bergman*. I have enjoyed our scientific and non-scientific discussions both inside and outside the walls of the workplace. And *Heidi*, a special thank you for boosting my ego when I need it and for selflessly sharing your ideas, feelings and adventures to always make me laugh so hard I cry. My dear "work husband" *Max Lönnfors* deserves a lot of thanks for all the more or less serious discussions, the endless support and hilarious travel-company. Thank you for taking care of me when I need it and for being a true friend. There are a few loyal people in the world that you can always rely on. You are one of them!

Thank you also to my friends outside of academia for still remembering me even though I have been busy from time to time. I also want to thank my mom *Barbro*, my dad *Berndt* and my brother *Conny*. Mamma och pappa, tack för att ni alltid uppmuntrar mig att fatta egna beslut och för att ni stöder mig i mina val. Trots att ni kanske inte alltid förstår vad det är jag gör så tack för att ni intresserar er och tror på att jag klarar av det jag tar mig för. Tack för all hjälp, all uppmuntran och all tid ni lagt ner på att lyssna och stöda mig. Tack också till morfar *Svante* för att jag har fått vara i Bromarf och stänga ute resten av världen. Det har betytt mycket för att få det här projektet avslutat. I also want to thank *Fábio* for always being there for me and supporting me through everything. Thank you for always showing how proud you are of me and making me feel good about my accomplishments and myself. You are my best friend and I sincerely thank you for the amazing patience you have.

Thank you also to Åbo Akademi Graduate School for believing in my science and accepting me as one of the first members of the school. The financial support has been indispensable. I would also wish to thank all the generous funding from the Sigrid Juselius Foundation, Åbo Akademi, ISB, Medicinska Understödsföreningen Liv och Hälsa r.f., Tor, Joe och Pentti Borgs Minnesfond, Magnus Ehrnrooth Foundation and Svenska Kulturfonden.

Käthe Dahlström

Åbo, October 2015

Abbreviations

3D	Three-dimensional
AIDS	Acquired Immunodeficiency Syndrome
APC	Adenomatous Polyposis Coli
ArmRP	Armadillo Repeat Fold/Protein
BLAST	Basic Local Alignment Search Tool
BLOSUM	Block Substitution Matrix
CASP	Critical Assessment of Protein Structure Prediction
CIP2A	Cancerous Inhibitor of Protein Phosphatase 2A
CSI-BLAST	Context-Specific Iterated BLAST
DAPK1	Death-Associated Protein Kinase 1
EM	Electron Microscopy
FucO	*Escherichia coli* propanediol oxidoreductase
GenPept	GenBank Gene Products Database
GPCR	G Protein-Coupled Receptor
HIV	Human Immunodeficiency Virus
I-TASSER	Iterative Threading Assembly Refinement
KpLpxO	*Klebsiella pneumoniae* LpxO
LPS	Lipopolysaccharide
MD	Molecular Dynamics
mTORC1	Rapamycin Complex 1
NAD^+	Nicotinamide Adenine Dinucleotide
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
PAM	Point Accepted Mutation
PDB	Protein Data Bank
Phyre	Protein Homology/Analogy Recognition Engine
PIR	Protein Information Resource
PIR-PSD	Protein Information Resource Protein Sequence Database
PP2A	Protein Phosphatase 2A
PSI-BLAST	Position-Specific Iterated BLAST
PSSM	Position-Specific Scoring Matrices
RCSB	Research Collaboratory for Structural Bioinformatics
RefSeq	Reference Sequence Collection
RMSD	Root Mean Square Deviation
SCOP	Structural Classification of Proteins
SMART	Simple Modular Architecture Research Tool
StLpxR	*Salmonella typhimurium* LpxR
t^6A	N^6-threonylcarbamoyl Adenosine
TrEMBL	Translated European Molecular Biology Laboratory
UniParc	Universal Protein Resource Archive
UniProt	Universal Protein Resource
UniProtKB	Universal Protein Knowledgebase
UniRef	Universal Protein non-redundant Reference Database
YeLpxR	*Yersinia enterocolitica* LpxR

1 Introduction

In order to function properly, all living organisms are dependent upon the engines of life, *i.e.* proteins. Vital processes like hearing, vision, smell, metabolism, immune response and cell division, to name a few, all involve proteins and these macromolecules are also the targets of most current medicines. Today, it is widely accepted that proteins are composed of amino acids encoded by nucleotides in a gene, but this idea was first introduced only after the Second World War. The general acknowledgement that proteins carry information and are built from polypeptides of amino acids, whose sequence specify the three-dimensional (3D) structure the protein folds into, was greatly influenced by Frederick Sanger and colleagues at Cambridge University, who sequenced the first complete protein, insulin, and Christian Anfinsen and colleagues at the National Institute of Health, who denatured ribonuclease and showed that it spontaneously refolds and regains enzymatic activity (Anfinsen, 1973; Hagen, 2000; Sanger, 1959). In the late 1960s, Pehr Edman gave a major push to the sequencing of proteins with his sequencing machine, which automated protein sequencing and made it routine work (Edman & Begg, 1967). This led to an increase in the known amino acid sequences, which in turn raised the need for collecting and storing the sequence data to make it available and accessible to all interested researchers. This was the birth of sequence databases (Hagen, 2000).

Margaret Dayhoff became one of the key contributors to the formation of the sequence databases during this era by cataloguing all available amino acid sequences in the Atlas of Protein Sequence and Structure (Dayhoff & Eck, 1968; Dayhoff, 1978). This annual publication eventually turned into the Protein Information Resource (PIR), a major online database established in 1983 (Hagen, 2000). The increased number of known amino acid sequences showed that some of them were more alike than others, *i.e.* related or homologous, and great efforts were made to develop computer algorithms to compare sequences and determine homology. This resulted in programs for sequence alignment, which is now an integral method for computational studies of proteins. Even more computationally challenging problems like protein structure modeling have their predecessors in the 1960s. In that decade, Cyrus Levinthal and researchers modeled the 3D structure of cytochrome c, but it was not considered to be a great breakthrough due to slow and less evolved computers (Hagen, 2000; Levinthal, 1966). Today, however, this early stage modeling can be seen as an important historical bridge to the advanced computer models of today.

Nowadays, the term bioinformatics, which became commonplace in the late 1980s, is divided into two main categories: sequence- and structure-based bioinformatics (Choong *et al.*, 2013; Hogeweg, 2011). Genome analysis,

sequence alignment, networks and evolution are all examples of bioinformatics with focus on sequences, while structural bioinformatics encompasses the prediction of the protein 3D structure and the structure-function relationship. This thesis focuses on the latter one of these, *i.e.* structural bioinformatics applications, but it also deals with alignments and sequence analysis, since these are an integral part of protein structure prediction. Although structural bioinformatics is more than 50 years old, it continues to be one of the most active areas of all bioinformatics research and the number of available tools and webservers may seem like a jungle for the user (Zhang, 2008a). Therefore, this thesis aims to guide the reader through the protein structure prediction process and to point out the facts that should be taken into account to avoid errors in the final structure. Several case studies are presented to illustrate specific scientific questions, where the predicted structures have provided useful information.

2 Review of the literature

The human genome was sequenced in 2001 and led to estimations that there are around 25,000 protein-encoding genes (Lander *et al.*, 2001; Venter *et al.*, 2001). Proteins are composed of 20 different, naturally occurring amino acids and their combination forms the primary structure, which is the unique linear peptide chain for each protein. Peptide bonds are formed through a reaction of the carboxyl group in one amino acid with the amino group of another amino acid and concomitant release of water. This reaction couples the amino acids together and starts the polypeptide from the N-terminus and continues it towards the C-terminus. The carbonyl groups in the peptide chain form a hydrogen-bonding pattern with the hydrogen atoms of the amino groups and, thereby, create regular and stable secondary structures called α-helices and β-strands (Pauling *et al.*, 1951). Moreover, multiple β-strands usually come together and form β-sheets (Pauling & Corey, 1951). In addition, irregular loops and turns link the stable and regular secondary structures together in a specific 3D pattern in space, called the tertiary structure. Hence, the unique sequence of amino acids in the primary structure determines the folding of the protein into its tertiary structure, also called the native or functional structure since the sequence-protein-structure paradigm states that the biological function of a protein is dependent on the right 3D fold of the protein (Anfinsen, 1973). One of the main driving forces behind protein folding is the hydrophobic effect, which makes the hydrophobic amino acids cluster together in the protein interior, while polar and charged amino acids are on the surface to interact with the surrounding water (Kauzmann, 1959; Lesk, 2000). Hydrogen bonds, together with hydrophilic and hydrophobic interactions, further stabilize the tertiary structure (Gibas & Jambeck, 2001; Lesk, 2002; Nelson & Cox, 2005). Some proteins are formed by more than one subunit and, therefore, they have a quaternary structure determined by the specific 3D arrangement of two or more monomers (Lesk, 2002; Nelson & Cox, 2005).

2.1 Determination of protein 3D structure

2.1.1 Experimental determination of protein 3D structure

There are three main methods for solving a protein structure experimentally: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (EM). When determining a protein structure by X-ray crystallography, the purified protein is crystallized and the crystal is subjected to an intense synchrotron X-ray beam. The electrons in the protein scatter the X-rays in a specific pattern, which is then used to calculate an electron density map where the amino acids in the protein are fitted to give a structure (McPherson, 2004). In NMR spectroscopy, the solution of the

purified protein is placed in a strong magnetic field and then probed with radio waves (Marion, 2013). The resonance can be observed in a spectrum, enabling determination of which atom nuclei are close to each other and analysis of the atom bond conformation. These restraints are then used to build the structure of the protein. On the other hand, EM is used for large macromolecular complexes, which are subjected to a beam of electrons to obtain a 3D image (Kuhlbrandt, 2013). EM has often been used in combination with X-ray crystallography or NMR spectroscopy to obtain the atomic details of the complex, but now EM alone has become a force to be reckoned with (Callaway, 2015). It can produce high-resolution models quickly, also from molecules that X-ray crystallography and other approaches have not been able to solve. However, X-ray crystallography relies on obtaining a protein crystal, while NMR spectroscopy is limited to low molecular weight proteins and, in addition, all the experimental methods are tedious, time-consuming and expensive. Therefore, structural modeling of proteins with the help of computers has caught increasing interest during the last decades.

2.1.2 Computational modeling of protein 3D structure

Structural bioinformatics is the branch of bioinformatics, which focuses on the prediction of 3D macromolecular structures, such as protein 3D structure (Altman & Dugan, 2003; Zhang *et al.*, 2005). One of the main questions, *i.e.* the protein structure prediction problem, concerns the challenge to understand how the information from the protein primary structure is translated into a 3D structure, and how to use this information for development of computational 3D structure prediction methodologies (Creighton, 1990). Several algorithms and methods have been established to solve this problem: homology modeling (also called comparative modeling), fold recognition methods, and first principle prediction with and without database information are all examples of these (Floudas *et al.*, 2006).

2.1.3 Template-free methods

First principle prediction without database information is also known as the *ab initio* method and considers only the amino acid sequence in search of the protein 3D structure that corresponds to the global free energy minimum (Bonneau & Baker, 2001; Osguthorpe, 2000). *Ab initio* methods are limited by the many possible conformations that the polypeptide chain can adopt and that represent local minima, but the global minimum is the structure of interest (Dorn *et al.*, 2014). However, the advantage of these methods is the ability to predict new folds since there is no requirement for templates with known structures. The other template-free method, first principle prediction with database information, uses sub-sequences of the protein of interest (target) to scan protein databases for general folding rules of similar

fragments, which are then assembled into a low energy structure with the help of scoring functions and optimization (Floudas *et al.*, 2006). Likewise, challenges with this method are mostly related to the many conformations the structure can adopt, but also to minimize the energy in the regions where the fragments are combined (Dorn *et al.*, 2014). This method shows, however, a clear advantage over *ab initio* methods in the sense that the conformational search space is more limited due to the use of similar fragments. An additional advantage is the possibility to predict new folds, similarly to *ab initio* methods.

2.1.4 Template-based methods

Template-based methods like homology modeling and threading use previously known protein structures (templates) and have proven themselves useful by producing better and more accurate protein structure models than the template-free methods mentioned earlier (Mullins, 2012). They also have the distinct advantage of being able to predict the structure of longer protein sequences, especially by combining multiple templates. Threading is a fold recognition method based on the conclusion that structure is more conserved than sequence and, therefore, two proteins can have the same fold although there is no apparent sequence similarity and evolutionary relationship between them (Finkelstein & Ptitsyn, 1987; Floudas *et al.*, 2006; Levitt & Chothia, 1976; Setubal & Meidanis, 1997). Here, the question of interest is merely whether the target protein can be reasonably represented by a known protein structure and, hence, modeled based on it (Tramontano, 2006). The threading process proceeds by placing the target protein sequence sequentially onto the known 3D structure in an optimal way and, through this, identifying homologous (evolutionary related) or analogous (no direct evolutionary relationship) templates (Dorn *et al.*, 2014). The energy of the target sequence in a certain 3D fold assesses the quality and is used to estimate the likelihood of the query sequence to adopt this particular fold. On the other hand, homology modeling relies on the evolutionary relationship between proteins. Homologous proteins are related to each other through a common ancestral protein, but their evolution has followed different paths and caused them to change. Homologs can be further divided into orthologs and paralogs, of which orthologs have evolved independently in different species but the function is similar. Meanwhile, paralogs are found in one species but the proteins have acquired different functions. Therefore, homology modeling is based on the assumption that similar protein sequences, *i.e.* homologous proteins, fold into a similar 3D structure. This method uses the target protein sequence and aligns it against the sequence of a homologous protein with known structure, which provides the modeling process with the structural information needed (Blundell *et al.*, 1987; Johnson *et al.*, 1994; Sali, 1995; Sanchez & Sali, 1997). This method can be used

whenever there is detectable similarity between the target protein and the template sequences.

2.2 Homology modeling

The first step in homology modeling is to retrieve the target protein sequence and use it to identify a homologous protein with known structure. The sequence of the template structure is aligned to the target sequence, whose 3D structure is thereafter modeled using the 3D coordinates of the known structure. The resulting structural model of the target protein is then validated and assessed. The steps from sequence alignment and onwards, are iterated until an acceptable structure for the target protein is aquired.

2.2.1 Databases

Major improvements in DNA sequencing techniques have enabled large-scale sequencing projects, which have increased the number of new protein sequences (Pavlopoulou & Michalopoulos, 2011; UniProt Consortium, 2015). Searching for a homologous sequence in several integrated protein sequence data repositories is usually the first step to characterize an unknown gene or protein. GenBank Gene Products Database (GenPept) at the National Center of Biotechnology Information (NCBI) (Wheeler *et al.*, 2003) contains amino acid sequences derived from translations of the corresponding nucleotide sequences, the entries have minimal annotation and several records can represent one protein. NCBI's Entrez Protein is similar, but adds additional information to the entries. NCBI also hosts the non-redundant Reference Sequence (RefSeq) collection, where the majority of the sequences are automatically generated but with only one record per protein. Moreover, there are universal curated databases with validated information in addition to the sequence data (Apweiler *et al.*, 2004). For example, Protein Information Resource Protein Sequence Database (PIR-PSD) has non-redundant entries organized in families and superfamilies with information about the protein itself, the function, the structure, bibliography and genetic data. For a long time, the leading curated protein sequence database was Swiss-Prot with non-redundant entries annotated by biologists, including information about function, post-translational modifications, domains, structure, diseases, location, pathways and variants. This type of manual annotation is tedious and, therefore, the Translated European Molecular Biology Laboratory (TrEMBL) database was developed in order to get the high number of new sequences accessible fast by having the entries computationally annotated.

The next generation of protein sequence databases takes one step further: the Universal Protein Resource (UniProt) incorporates Swiss-Prot, TrEMBL and PIR-PSD into a single resource and provides researchers with

comprehensive, high-quality and freely accessible protein sequence and function information (UniProt Consortium, 2015). This leading protein sequence database is built from three components: the UniProt Archive (UniParc), the UniProt non-redundant reference databases (UniRef) and the UniProt Knowledgebase (UniProtKB). UniParc is the most comprehensive non-redundant sequence collection with publicly available protein sequences from Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl, International Protein Index, Protein Data Bank (PDB), RefSeq, FlyBase, WormBase, and the patent offices in Europe, the United States and Japan. The current release statistics show 92,444,468 entries (UniProt Consortium release 2015_05, 29.4.2015). In turn, UniRef is a non-redundant sequence collection clustered by sequence identity and taxonomy with 59,744,893 entries (UniProt Consortium release 2015_05, 29.4.2015). Moreover, UniProtKB is a merger of Swiss-Prot, TrEMBL and PIR-PSD and is the central annotated database for information on protein sequence and function. It contains a reviewed, manually annotated section (UniProtKB/SwissProt) with 548,454 entries (UniProt Consortium release 2015_05, 29.4.2015), as well as a section without review but instead automatically annotated records (UniProtKB/TrEMBL) containing 47,452,313 entries (Figure 1 a and b). The number of deposited sequences has increased almost exponentially for both sections, but around 2010 the UniProtKB/SwissProt curve reaches a plateau (Figure 1a). This effect can partly be accounted for by redundancy of the new sequences, *i.e.* they are splice variants or mutants of already annotated sequences, which means that the information is incorporated into an existing entry. Also the number of entries in UniProtKB/TrEMBL has made a sudden drop in 2015 due to a new procedure to identify highly redundant proteomes, which has been applied to bacterial proteomes and redundant sequences have been removed (Figure 1b).

Also experimentally determined protein structures are deposited in a database. PDB is the single worldwide, publicly available repository of free of charge macromolecular structure data with the aim to facilitate the use and analysis of structural data, thereby enabling new science (Berman *et al.*, 2000; Berman *et al.*, 2002). PDB was first established at Brookhaven National Laboratory (Bernstein *et al.*, 1977) in 1971 with seven structures, but the 1990s saw a dramatic increase in the amount of deposited structures (Figure 1c) due to improvements in the methods for solving protein structures (Berman *et al.*, 2002). Since 1998, the Research Collaboratory for Structural Bioinformatics (RCSB), *i.e.* Rutgers, the State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the National Institute of Standards and Technology, has managed PDB. The content of data in the PDB includes, for example, the source of the protein, the sequence, method for solving the structure, possible ligands and cofactors, 3D coordinates for the structure, and literature citations (Berman *et*

al., 2002). PDB contains 101,233 (19.05.2015) known protein structures, which is remarkably less than the available sequences in the wider UniProt/TrEMBL database (47,452,313 entries) (compare Figure 1 b and c). Moreover, some structures in PDB are partially solved, *i.e.* only a single domain of the protein is known, and the database is also redundant with many different variants of one protein due to mutations or complex structures with a ligand, co-factor or inhibitor. When adjusting for 95 % sequence identity, the number of available structures drops to 43,580, which means that 8 % of the sequences in the smaller UniProt/SwissProt database (548,454) are structurally characterized. When compared to the wider UniProt/TrEMBL database, the same value becomes 0.09 %. Hence, there is a wide gap between known sequences and known structures.

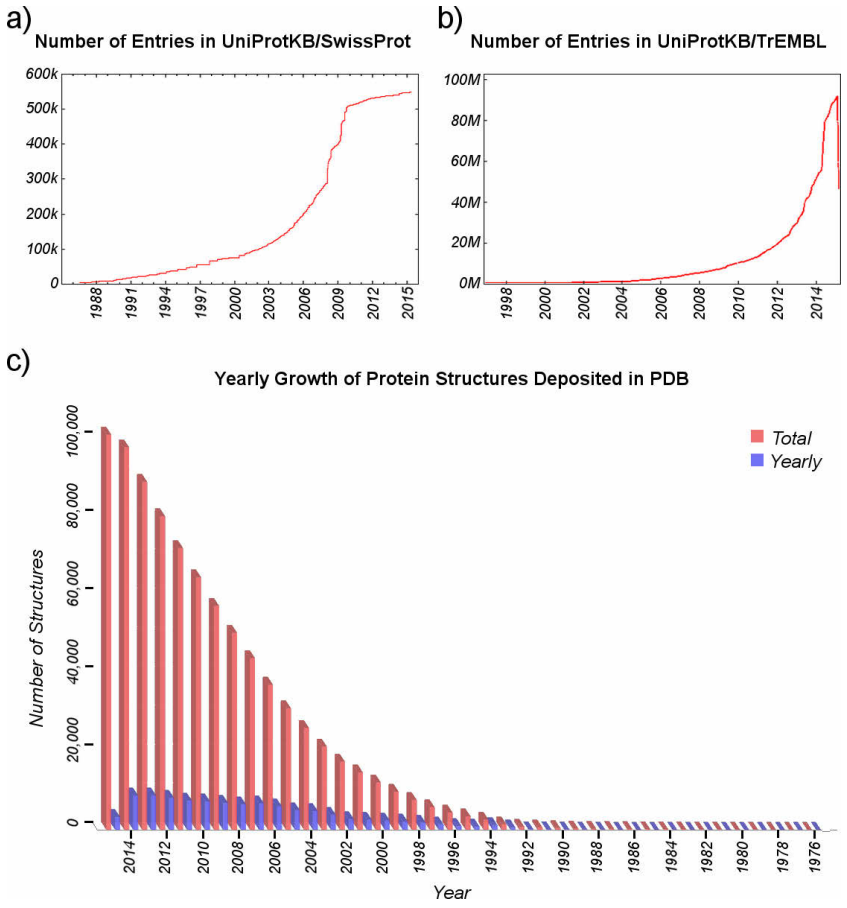


Figure 1. Database growth. a) the current number of entries (548,454) in UniProtKB/SwissProt, which is reviewed and manually annotated. b) the number of entries (47,452,313) in UniProtKB/TrEMBL, which contains automatically annotated records without review. c) the number of crystallized structures (101,233) deposited in the Protein Data Bank (PDB).

Grouping of the proteins in PDB according to structural similarity has also resulted in databases such as the manually annotated Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995), which is now updated to SCOP2 (Andreeva *et al.*, 2014) and the manually and automatically classified CATH (Sillitoe *et al.*, 2015). According to SCOP (v.1.75), there are 1393 unique folds for the proteins in PDB since 2008, while CATH (v.4.0.0) has defined 1375 unique folds since 2012, indicating that the majority of the structures being solved have a similar fold to the already known ones. Hence, a small number of unique folds represent the majority of known structures (Orengo & Thornton, 2005).

2.2.2 Prediction of protein primary structure

After retrieving the target sequence it can be used for several analyses. The first step is to study the domain composition of the protein and, if different domains exist, identify the types of domains. Domains are usually compact areas, which fold independently, are spatially separated from each other and may have a defined semi-independent function. Therefore, the combination and cooperation of different domains give rise to proteins with complex or multiple functions (Holland *et al.*, 2006; Ponting & Russell, 2002). There are various databases, for example Simple Modular Architecture Research Tool (SMART) (Letunic *et al.*, 2015; Schultz *et al.*, 1998) and Pfam (Finn *et al.*, 2014), which are dedicated to finding domains in query proteins. The different protein domains can give valuable information about the function of an uncharacterized protein if the domain is found in so called signature databases. These are databases of consensus repeats crucial for the structure or function of domains or protein families (Mulder & Apweiler, 2002; Wu *et al.*, 2003). The signature consists of different levels: motifs, fingerprints, patterns and profiles. Motif is the smallest constituent with typically 10-20 amino acids forming a single conserved region. On the other hand, a group of several motifs are called a fingerprint, while a pattern highlights a consensus sequence, *i.e.* specific amino acids at a certain position and in a unique order. Patterns are usually short and restricted to the most conserved regions in the protein sequence (Hofmann, 2000). Furthermore, profiles are used to describe and detect larger areas or domains of the sequence, including variable regions (Gribskov *et al.*, 1987) (Figure 2). A widely used server for detecting sequence patterns and profiles is PROSITE (de Castro *et al.*, 2006).

2.2.3 Prediction of protein secondary structure

The earliest methods for secondary structure prediction were based on the probability of a certain amino acid to be in a specific secondary structure. For example leucine, isoleucine and valine are common in β-strands (Chou & Fasman, 1974). Today, the secondary structure prediction accuracy has improved to over 70 %, in a large part due to incorporation of multiple

sequence alignment information. This enables identification of highly conserved regions in the protein sequences, which translates into structurally or functionally important regions that are mostly concentrated to secondary structure elements in the protein, while variable regions make up loops on the surface of the protein.

The assumption that proteins with > 30 % sequence identity have a similar fold allows the secondary structure prediction programs to incorporate information from proteins with known structure, as well as multiple sequence alignments and sequence profiles to deduce evolutionary relationships (Pavlopoulou & Michalopoulos, 2011; Rost, 1999). Examples of secondary structure prediction servers are PSIPRED with an accuracy of 76.3 % (Jones, 1999; McGuffin *et al.*, 2000) and PORTER, which has an accuracy of 82.2 % (Mirabello & Pollastri, 2013; Pollastri & McLysaght, 2005). Furthermore, JPred has an accuracy of 81.5 % (Cole *et al.*, 2008), which is a result from implementation of the knowledge that core amino acids play an important role for protein folding since they need to be buried within the protein (Chan & Dill, 1990). Hence, the accuracy of the secondary structure prediction is improved by measuring how accessible a residue is to the solvent (Adamczak *et al.*, 2005).

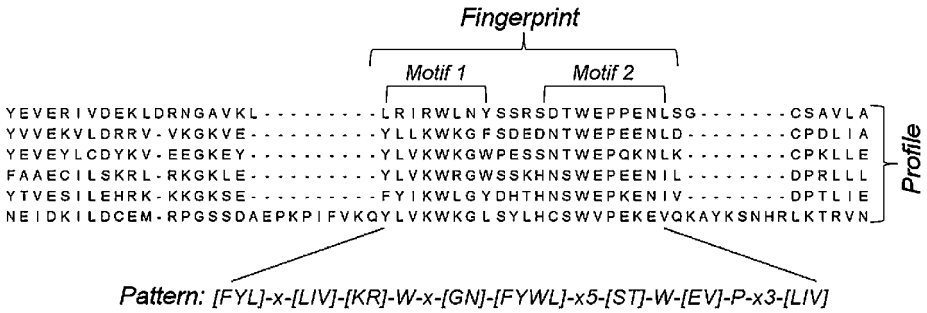


Figure 2. Protein signatures. Motif is the smallest constituent and several motifs form a fingerprint. A pattern highlights a short consensus sequence, *i.e.* specific amino acids at certain positions and in a unique order. Profiles are used to describe and detect larger areas or domains of the sequence, including variable regions. Figure adapted from Pavlopoulou & Michalopoulos, 2011.

2.2.4 Transmembrane proteins

Transmembrane proteins carry out cell signaling events, molecular transport and many other important biological functions (Schulz, 2002; von Heijne, 1996). These are proteins, which span the lipid membrane surrounding an

organelle and they are classified as either α-helical transmembrane proteins or transmembrane β-barrel proteins. The classification depends on the structure of the membrane-spanning segment: α-helical transmembrane proteins are anchored to the lipid bilayer by one or more α-helices, while transmembrane β-barrel proteins form a barrel-like channel with their membrane-spanning antiparallel β-strands (Schulz, 2000). The structure of transmembrane proteins is difficult to determine experimentally, which makes computational tools essential for prediction of the presence and topology of transmembrane proteins and, ultimately, a possible function (Sonnhammer *et al.*, 1998). These programs mostly base their prediction on the amino acid sequence of a protein, where for example a stretch of 15-30 hydrophobic amino acids indicates a transmembrane α-helix (Schulz, 2002). Furthermore, the topology can be predicted by taking into account the positive inside rule, which states that positively charged residues are more predominant in loops on the cytoplasmic side of the membrane (von Heijne, 1992). Transmembrane α-helices can be predicted with the TMHMM server (www.cbs.dtu.dk/services/TMHMM/) and TMpred (Hofmann & Stoffel, 1993). Transmembrane β-barrel proteins are not as easily predicted because of the shortness of the transmembrane segments, as well as an uncertain distribution of polar and non-polar amino acids (Schulz, 2000). Despite this, there are available computational techniques like TBBpred with an accuracy of 81.8 % (Natt *et al.*, 2004) and BOCTOPUS with an accuracy of 87 % (Hayat & Elofsson, 2012), which are dedicated specifically to the prediction of transmembrane β-barrel proteins (Pavlopoulou & Michalopoulos, 2011).

2.2.5 Sequence and structure searching

Databases contain thousands of sequences or structures, which need to be scanned to find homologs. The Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) at NCBI enables scanning of a nucleotide (BLASTN) or protein sequence (BLASTP) (Altschul & Koonin, 1998) against a selected database in search of local similarity, which is reported as a statistical significance. There are different variants of BLAST, like the position-specific iterated BLAST (PSI-BLAST) (Altschul *et al.*, 1997) and context-specific iterated BLAST (CSI-BLAST) (Biegert & Söding, 2009). PSI-BLAST uses position-specific scoring matrices (PSSM), which represent multiple sequence alignments with numbers so that each number indicates the probability of a certain amino acid at every position. The number is also affected depending on whether the substitution is conserved or not. When searching a database with PSI-BLAST, the target sequence is first used for standard BLASTP search and the statistically significant results are aligned together with the target sequence to create the PSSM. This matrix representing the collective characteristics is then used to search the database for more related sequences, which is why PSI-BLAST is able to detect distantly related proteins that may not be found with regular BLASTP. While

PSI-BLAST considers each position in the sequence on its own, CSI-BLAST considers each amino acid position centered between twelve surrounding positions, which means that the context the amino acid is in affects the generation of the PSSM. This matrix is then used in the PSI-BLAST search to find more related sequences. It was shown that CSI-BLAST can detect ~15 % more homologs than PSI-BLAST. However, PSI-BLAST and CSI-BLAST pose the possible problem of having a higher rate of false positives, *i.e.* sequences that are unrelated to the query sequence (Pavlopoulou & Michalopoulos, 2011).

2.2.6 Structure analysis

Before choosing a known protein structure to be used as a template, the structural data should be carefully analyzed. Special attention should be paid to the target – template sequence identity, and the higher the overall sequence identity is, the better (Tramontano & Morea, 2003). However, if there are many possible templates, other factors can play a key role in the analysis: presence or absence of cofactors and ligands, oligomerization state and conformation, to name a few (Kopp & Schwede, 2004). Furthermore, the quality or resolution of the crystal structure is essential, since the performance of the modeling programs is highly dependent on the quality of the input data, which will directly affect the quality of the model in the end. A low-resolution crystal structure can contain errors in areas without clearly defined electron density, which means that a higher resolution crystal structure is a more accurate template for modeling. Also, is it of interest to study the target protein in the presence of cofactors and/or ligands? Does the target protein change conformation or oligomerize? If a ligand bound state or a specific conformation is the main aim of the modeling, then these rise as important criteria and affect the selection of a proper template.

Protein structure comparison

Protein structures are often compared to each other as well to highlight the similarities and differences. Structural comparison also helps to infer evolutionary relationship even when the proteins have less than 25 % sequence identity (Johnson & Lehtonen, 2000; Laurents *et al.*, 1994) and it is also used for classification of proteins and their domains into families (Murzin *et al.*, 1995; Orengo *et al.*, 1997). The protein structure comparison is done by superimposition of two or more structures and, during the process; one of the molecules is rotated and oriented to fit on top of the other molecule (Maiti *et al.*, 2004). The simplest way of doing this is to find a set of reference points, which need to have maximal correspondence when the structures are superimposed. When comparing protein structures, the question of interest is how similar the structures are, whether it is at a local level around the ligand-binding site or if it is the global fold. The most commonly used measure for this is the root mean square deviation (RMSD), which is

calculated by adding together the square of the difference in distance (Ångström [Å]) between equal C^{α}-atom pairs and dividing the sum by the number of compared atoms. Hence, the lower the RMSD, the more similar are the compared structures.

2.2.7 Alignment

During evolution, amino acid sequences of related proteins diverge and acquire mutations, insertions or deletions of amino acids. The changes in sequence form a pattern of importance, where highly conserved areas are presumed to be structurally or functionally important for the protein and are detectable through pairwise or multiple sequence alignment of homologous sequences. When aligning sequences, the purpose is to maximize the number of aligned identical amino acids and keep the gaps caused by insertions and deletions to a minimum with the help of a gap penalty. The accuracy of a homology model is highly dependent on the alignment accuracy, which is why a high sequence identity between the aligned sequences is preferable (Mullins, 2012). For example, when the sequence identity drops below 30 %, errors in the alignment are more prone to occur because the amino acid sequences have diverged too much for the sequence alignment to be accurate and reliable, thereby lowering the accuracy and reliability of the resulting model. However, the three-dimensional structure of a protein is more conserved than the sequence, which makes a structure-based alignment more advantageous. To obtain the structure-based alignment, the protein structures are superimposed and amino acids in corresponding three-dimensional positions are aligned with each other in the alignment.

Pairwise sequence alignment

Pairwise sequence alignment is mainly used for finding homologous proteins in sequence or structure databases and to align two very closely related sequences. There are two ways to perform a pairwise sequence alignment: global and local sequence alignment. The global alternative tries to find the optimal alignment for the entire length of the two sequences based on the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970). This algorithm handles the sequence boundaries as edges and to create the optimal global alignment, the search path must start at one edge and reach the other edge. A version of the Needleman-Wunsch algorithm is the Fredman algorithm, which follows the same approach but the execution is faster (Fredman, 1984). On the other hand, local sequence alignment has implemented the Smith-Waterman algorithm (Smith & Waterman, 1981) to identify local regions with similarity within two sequences lacking relevant similarity over their entire length. Hence, this algorithm allows the edges to start and end within the sequence and the most similar substrings from both sequences are aligned.

Multiple sequence alignment

If the target and the template proteins share a low sequence identity and pairwise sequence alignment is unreliable, the accuracy and the reliability of the alignment can be improved by using multiple sequences in the alignment. By choosing several sequences within a sequence identity range for both the target protein and the template protein, the evolutionary gap between these two sequences is lowered and, hence, sequences are more likely to be correctly aligned. Moreover, a multiple sequence alignment can reveal conserved amino acid patterns, which would not be distinguishable in a pairwise alignment. The essence of this can be summarized in the following quote from Hubbard *et al.*, 1996: "one or two homologous sequences whisper about their three-dimensional structure; a full multiple alignment shouts out loud". Usually, a multiple sequence alignment is done progressively by first creating a crude tree to determine the relationship between the sequences and use it as a base for the order in which the sequences are added to the alignment (Feng & Doolittle, 1987). The closely related sequences are aligned first and, thereafter, the more distant ones are added to the alignment according to the predetermined order. CLUSTALW (Larkin *et al.*, 2007) and T-Coffee (Notredame *et al.*, 2000) are examples or popular programs for multiple sequence alignment (Daugelaite *et al.*, 2013; Wallace *et al.*, 2006). In this work, multiple sequence alignments were produced with the program MALIGN (Johnson & Overington, 1993) in the BODIL modeling environment (Lehtonen *et al.*, 2004). MALIGN compares the sequences based on the dynamic programming algorithm of Fredman and it also constructs multiple sequence alignments according to the Feng & Doolittle approach described above (Johnson *et al.*, 1993).

Structure-based alignment

Despite their usefulness, multiple sequence alignments might not always return the most accurate results and there is a need for improvement, especially when it comes to large-scale analysis and low sequence identity (Armougom *et al.*, 2006). A more reliable approach is to combine structure alignment with sequence alignment (Holm & Sanders, 1996). This is a direct effect of the structure stability in evolution, *i.e.* structures are more conserved than sequences, and results in constraints on the alignment. However, structure-based alignments require known protein structures, but the more structural data that is made available, the better the structure-based sequence alignments can perform. This type of method is applied by the T-Coffee variant 3D-Coffee (O'Sullivan *et al.*, 2004). Moreover, it is possible to superimpose structures on top of each other and generate an alignment of amino acids in the same 3D positions, to which the target protein is then aligned. The structure-based alignment can for example be done with the program VERTAA (Johnson & Lehtonen, 2000) in the BODIL modeling environment (Lehtonen *et al.*, 2004). VERTAA calculates the initial set of

common elements in two compared structures by correlating the number of C^{α}-atoms within 14 Å of each residue for the whole sequence of each structure and then superimposing them. Thereafter, the target protein can be aligned with MALIGN (Johnson & Overington, 1993) using the pre-aligned function to keep the structure-based alignment static. Hence, the structural information is taken into account in the alignment process and the spatial variations are introduced into the gap penalties, which ultimately leads to more accurate alignments.

Amino acid substitution models

Amino acids change during evolution but the probability for one amino acid to substitute another is not the same for all 20 amino acids. Instead, one amino acid is more prone to change to another amino acid with similar physicochemical properties, which means that the changes do not markedly affect the overall structure or function of the protein (Pavlopoulou & Michalopoulos, 2011). Hence, in addition to identical amino acids in aligned poitions, these conserved substitutions should also be considered. This is achieved through substitution matrices like the Point Accepted Mutation (PAM) matrix (Dayhoff *et al.*, 1978) and the BLOck Substitution Matrix (BLOSUM) series (Henikoff & Henikoff, 1992). The PAM matrices were developed from global alignment of closely related protein sequences (> 85 % identity) to study the substitution patterns (Dayhoff *et al.*, 1978). The results were described in tables, which indicate the frequency of two amino acids replacing each other at a specific position. One PAM unit (PAM1) equals one amino acid change per 100 amino acids, *i.e.* 1 % divergence. A series of PAM matrices has then been established by multiplying the PAM1 matrix by itself and the higher the number of the matrix, the more suited it is to align distantly related protein sequences. For example, multiplying the PAM1 matrix by itself 250 times generates the PAM250 matrix. BLOSUM matrices, on the other hand, are calculated from local alignments of functionally important and highly conserved blocks in homologous sequences. BLOSUM matrices also form a series, where the number indicates the identity percentage between the sequences used to develop the matrix. For example, the BLOSUM80 matrix is developed from substitutions in alignments of proteins sharing 80 % identity, while BLOSUM45 is based on a 45 % sequence identity between the aligned sequences. Therefore, in the case of BLOSUM matrices, the higher the number of the matrix, the more suitable it is to align closely related protein sequences. Another approach has been taken to develop the STRMAT110 matrix implemented in MALIGN (Johnson & Overington, 1993) in the BODIL modeling environment (Lehtonen *et al.*, 2004). This amino acid substitution matrix is built from the comparison of protein 3D structure through multiple structure alignment for each of 65 protein families and incorporation of information from both highly similar areas and more variable regions with gaps (Johnson & Overington,

1993). The STRMAT110 takes into account the local environment in the folded protein and also functional aspects, which together constrain the amino acid substitutions, insertions and deletions to various degrees (Johnson *et al.*, 1996).

Gap penalties
In addition to these substitution matrices, gap penalties are also considered when aligning sequences. Gaps in one sequence compensate for insertions in the other sequence and are usually introduced to improve the alignment. The number of gaps in the alignment needs to be reasonable to reflect a possible biological scenario (Baxevanis & Ouellette, 2004; Giribet & Wheeler, 1999). Gaps can be scored according to the affine gap penalty method, which deduces a fixed penalty score when a gap is introduced and then has a lower score for extending an already introduced gap and that is proportional to the length of the gap. However, most programs allow manual adjustments of these penalty scores. Another method is the non-affine, linear gap penalty method, which penalizes each gap position equally (Baxevanis & Ouellette, 2004; Giribet & Wheeler, 1999). Affine gap penalties consider the fact that one mutation event can insert or delete more than a single residue and also that, usually, conserved regions do not contain gaps, which is why this method enables detection of more distant homologs and also represents the biology in a more accurate fashion (Baxevanis & Ouellette, 2004).

These matrices and gap penalties are incorporated into all software, which often provide a default matrix and gap penalty. However, these might not always be the best ones for the question to be answered. Therefore, the user should consider the different options separately for each research question of interest. There are also specialized matrices for specific species, particular proteins etc. (Wheeler, 2002), which means that no single matrix is good enough to produce significant results for all biological questions.

2.2.8 Model building

Once a suitable template structure has been chosen and optimally aligned to the target protein, the model can be built. The coordinates of the template backbone are copied to the model, while the conformation of individual amino acid side chains has to be predicted by the model-building program, which employ database searches for optimal side-chain packing (Flohil *et al.*, 2002). The model building is challenged by insertions and deletions in the target protein. These are usually located to loops and in the case of insertions there is no template. To tackle this, the modeling programs can search structural libraries or use *ab initio* methods to find an energetically favorable loop conformation, especially when building longer loops (Fiser *et al.*, 2000). The last step is to relax and refine the model by energy minimization, release

of conformational strain and optimization of stereochemistry to remove unfavorable contacts (Mullins, 2012).

The different methods for model building are: rigid-body assembly, segment matching or modeling by satisfaction of spatial restraints. Rigid-body assembly is the oldest method and assembles a model from rigid fragments corresponding to core aligned regions (Blundell *et al.*, 1987; Greer, 1990). The rigid fragments are placed onto the template backbone, after which variable parts such as loops and side chains are rebuilt. On the other hand, the segment-matching approach employs a subset of atomic positions derived from the alignment to search databases for matching segments (Claessens *et al.*, 1989; Jones & Thirup, 1986; Levitt, 1992). In turn, the modeling by satisfaction of spatial restraints approach derives the restraints from the alignment and the model is built in such a way that these restraints are violated as little as possible (Sali *et al.*, 1990). Wallner & Elofsson (2005) showed that none of these methods significantly outperforms the others, but rather there are pros and cons with each of them. However, the three best performing modeling programs were MODELLER (Sali & Blundell, 1993), Nest (Petrey *et al.*, 2003) and Segmod/ENCAD (Levitt, 1992). MODELLER represents modeling by satisfaction of spatial restraints, Nest is a rigid-body assembly method and SegMod/ENCAD exemplifies the segment-matching approach. All of these produce chemically correct models within a reasonable time limit (Wallner & Elofsson, 2005). MODELLER (Sali & Blundell, 1993) is still a popular program for homology modeling, but also ORCHESTRAR (Tripos International), Prime (Schrödinger, LLC), MOE (Chemical Computing Group, Inc.), Composer (Blundell *et al.*, 1988; Sutcliffe *et al.*, 1987a; Sutcliffe, Hayes *et al.*, 1987b) and Robetta (Kim *et al.*, 2004) are commonly used (Dolan *et al.*, 2012).

The ability of template-based modeling to counteract the gap between known sequences and structures, coupled to the well-defined steps in the protein structure modeling procedure, have sparked the development of automated servers and pipelines for modeling. This reduces the required expertise and makes the homology modeling methods available to a broader audience (Kopp & Schwede, 2004; Mullins, 2012). These types of servers were pioneered by SWISS-MODEL in 1996 (Guex & Peitsch, 1997; Peitsch, 1996; Schwede *et al.*, 2003) and they provide homology modeling on demand through the Internet. Nowadays, there are also so called meta-predictors, through which the researcher can obtain a predicted model from several automated servers and thereafter choose the best model. Other examples of automated modeling servers are the Protein Homology/analogy Recognition Engine (Phyre) (Kelley & Sternberg, 2009), HHPred (Remmert *et al.*, 2011; Söding, 2005; Söding *et al.*, 2005) and Iterative Threading ASSEmbly Refinement (I-TASSER) (Roy *et al.*, 2010; Roy *et al.*, 2012; Zhang, 2008b),

of which the last one was proven to produce the best 3D structure predictions among all automated servers in the Critical Assessment of Protein Structure Prediction (CASP) experiments 7-10 (Zhang, 2014). Whichever path is chosen for production of the model, whether it is performing each step in the homology modeling procedure separately or using fully automated servers, the resulting models have to be carefully examined and evaluated to assess the quality and accuracy.

2.2.9 Model evaluation and refinement

The quality and accuracy of the resulting models are assessed by the geometry of individual model regions and identification of possible errors. Depending on the evaluation results, the model can be used for various predictions and interpretations. Highly accurate models, which are based on a template protein with high sequence identity to the target protein, can be used for analysis of specific amino acids and their functional role for ligand-binding. If the template protein has low sequence identity to the target protein, it consequently follows that the model has lower accuracy and quality, which makes it more suited for general studies like determination of overall fold and amino acids contributing to active sites or ligand-binding. A commonly used evaluator of model accuracy is the RMSD value, which considers aligned residues and the distance between target and template α-carbon atoms. The common rule says that the lower the RMSD, the more accurate is the model (Pavlopoulou & Michalopoulos, 2011).

The likely overall quality of the resulting model can already be estimated at the alignment level, since the sequence identity will have the biggest effect on the quality of the final model (Mullins, 2012). If sequence identity between the target and the template protein is > 50 %, then the model will probably be of good accuracy with ~1 Å RMSD from the template (Chothia & Lesk, 1986). This corresponds to NMR structures of medium resolution or low-resolution X-ray structures (Read & Chavali, 2007). The side chains can show poor packing and the loops might need additional refinement, but the overall quality of the model is high. When the sequence identity is around 40 – 50 %, the model accuracy will be good with an RMSD < 2 Å, but even a model based on 30 – 40 % sequence identity can be significantly different in less accurate regions with over 2 Å RMSD from the known template structure and, therefore, it shows only medium accuracy. Low accuracy models are based on template proteins with < 30 % sequence identity, which causes the alignment errors to increase and the accuracy to decrease (Figure 3). Hence, it is possible to assess the model quality by considering the target – template sequence alignment and the probability or confidence that each pair of amino acids are aligned correctly (Chen & Kihara, 2008; Lassmann & Sonnhammer, 2005; Sadreyev & Grishin, 2004; Tress *et al.*, 2004).

Several model quality assessment programs have been developed to aid the critical ranking and selection of models. Model quality assessment methods normally predict global or local quality scores and they can be classified into single-model methods (Wang *et al.*, 2011; Yang & Zhou, 2008; Zemla *et al.*, 1999; Zemla *et al.*, 2001; Zemla, 2003) or multi-model methods (McGuffin, 2009; McGuffin & Roche, 2010; McGuffin *et al.*, 2013; Wang *et al.*, 2011; Wang *et al.*, 2011). Single-model methods try to predict the quality of a protein model based on the structural features (Kalman & Ben-Tal, 2010; Luthy *et al.*, 1992; Ray *et al.*, 2012; Tress *et al.*, 2003; Wallner & Elofsson, 2006). These methods are mostly based on physical effective energy terms from analysis of force fields or empirical pseudo energy derived from known protein structures, but they can also rely on the agreement between protein characteristics, such as secondary structure, solvent accessibility and contact maps (Pawlowski *et al.*, 2015). Examples of the single-model methods are the programs PROCHECK (Laskowski *et al.*, 1993), WHATCHECK (Hooft *et al.*, 1996), ProSA (Sippl, 1993; Wiederstein & Sippl, 2007), Verify-3D (Bowie *et al.*, 1991; Eisenberg *et al.*, 1997; Luthy *et al.*, 1992), ERRAT (Colovos & Yeates, 1993), QMEAN (Benkert *et al.*, 2009) and ProQ (Wallner & Elofsson, 2003). Both PROCHECK (Laskowski *et al.*, 1993) and WHATCHECK (Hooft *et al.*, 1996) base their assessment on stereochemical quality, such as main-chain bond lengths and bond angles. On the other hand, ProSA evaluates the model packing by estimating the probability for two residues to be at a specific distance from each other (Sippl, 1993; Wiederstein & Sippl, 2007). ProSA also takes into account the solvation of the model, *i.e.* the interactions between the model and the solvent. Verify-3D assigns an environmental class to each residue based on the secondary structure, the buried sections and the polar contacts (Bowie *et al.*, 1991; Eisenberg *et al.*, 1997; Luthy *et al.*, 1992). The probability of an amino acid to be in each type of environment is estimated and the sum of these probabilities indicates the model quality. In this case, the higher the probability, the more correct is the model. ERRAT is also based on the probability that two atoms of a particular type are in contact, but in this program, the fraction of all contacts of a particular type is used (Colovos & Yeates, 1993). The described programs have been developed to distinguish between native and non-native structures; however, model quality assessment programs have also been developed with the aim to find the best possible model. QMEAN is one of these programs and is based on six terms: local geometry assessed by torsion angles, two terms considering the distance between atoms, the burial of residues, and two terms describing the agreement between predicted and calculated secondary structure and solvent accessibility (Benkert *et al.*, 2009). ProQ is another program developed to find the best possible model (Wallner & Elofsson, 2003). This program is based on protein models of different similarity, which were each described by a set of structural features (atom-atom contacts, residue-residue contacts, surface area exposure, and secondary structure

agreement). These factors were then used to train a neural network to predict protein model quality. Overall, the main advantage of single-model methods is that they can address the important challenge in protein structure modeling – picking out a good model from the irrelevant ones (Wang & Cheng, 2012). On the other hand, multi-model methods generate a consensus or a cluster of models and compare the quality of one model to the other models in the pool (Kryshtafovych *et al.*, 2015). These methods are developed based on the assumption that frequently predicted conformations are likely to be closest to the native structure. Hence, a high quality score means that the model is similar to the rest of the models in the pool. These types of methods tend to work well when the models in the pool are of good quality and generated by multiple different protein structure prediction techniques. A known multi-model quality assessment program is Pcons, which was pioneering in this field (Wallner & Elofsson, 2006). Furthermore, new methods called hybrid quality assessment methods or quasi single-model quality assessment methods have been developed to benefit from the strengths of both single-model and multi-model methods (Cheng *et al.*, 2009). These methods use a single-model method to assess and score each of the input models and then compare them to a subset of previously generated models. An example of a quasi single-model quality assessment program is MODFOLD, which has been developed with the objective to find the best possible model (Buenavista *et al.*, 2012; McGuffin & Roche, 2010; McGuffin *et al.*, 2013).

The available modeling techniques and the quality assessment programs are also assessed themselves in the CASP experiments, where models are made for experimentally determined protein structures with unreleased structural data (Huang *et al.*, 2014; Moult *et al.*, 2014). Instead, researchers are allowed to model these specific proteins with various programs and servers in a blind fashion, which means that the assessors do not know the identity of the researcher. The resulting models are then compared to the experimentally determined protein structure and allows for analysis of the accuracy of specific modeling methods, resulting in a ranking of the predictor groups according to their success. CASP has highlighted the importance of the methods to evaluate the overall accuracy of the model, as well as the local accuracy at amino acid side chain level (Moult *et al.*, 2014). Moreover, in one of the latest CASP rounds, CASP 10, the results show that the selection of the most accurate model can be a difficult task and is an important area for development (Huang *et al.*, 2014). Nevertheless, the same CASP round saw a significant improvement in refinement of protein structure models with a concomitant improvement in accuracy (Moult *et al.*, 2014). Furthermore, the model accuracy estimations in CASP 11 actually showed that if models are available from more than one structure prediction server, or multiple models from one structure prediction server, the single-model or multi-model methods are equally good at choosing the best models (Kryshtafovych *et al.*,

2015). However, multi-model methods perform better when the aim is to identify worse models. Together with quasi-single methods, the multi-model methods are also better at estimating the local accuracy residue level. Overall, the individual accuracy of a model is the main determinant of how and what a model can be used for rather than the method used for modeling (Moult *et al.*, 2014). Therefore, each model should be critically assessed and analyzed. Naturally, the ultimate validation of a model comes from experiments (di Luccio & Koehl, 2011). Site-directed mutagenesis, cross-linking and mass spectrometry are just a few examples of experiments, which can contribute to the validation of a model. This type of experimental data can also be used as modeling constraints to improve the accuracy of the model. Overall, the evaluation of the modeling methods and the quality assessment programs themselves has resulted in a broad acceptance of models as recognized and well-accepted sources of structural information. Also, the accuracy of homology models has greatly improved due to better methods, larger amounts of sequences and structures in databases and evaluation by CASP (Moult *et al.*, 2014).

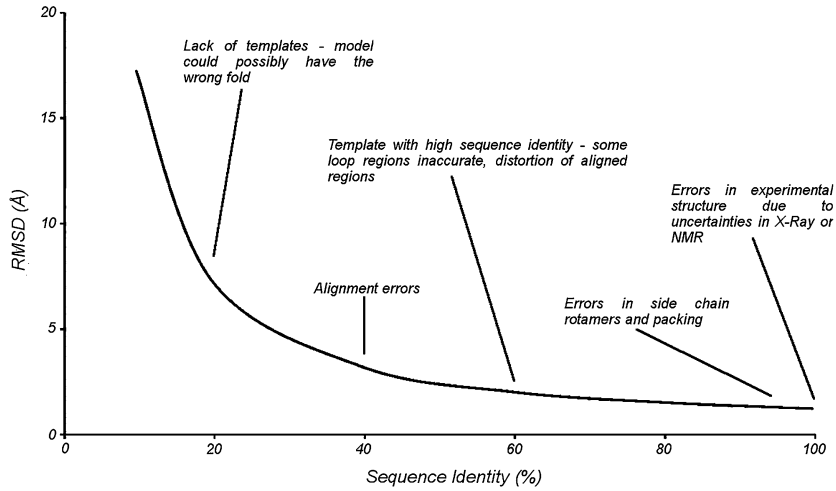


Figure 3. The relationship between sequence identity and the accuracy of the 3D structural model. Figure adapted from Mullins, 2012.

2.3 Use of predicted 3D structures

Homology modeling is able to predict protein 3D structures with high accuracy and is used for example in the pharmaceutical sector for drug design and virtual screening, as well as in molecular research, biotechnology and for

designing site directed mutagenesis studies (Hillisch *et al.*, 2004; Kopp & Schwede, 2004; Vangrevelinghe *et al.*, 2003). Homology modeling is also useful for constructing hypotheses about protein function or functional annotation (Hermann *et al.*, 2007), as well as for docking studies to analyze protein-ligand interaction patterns. This type of study can be used to identify critical areas for substrate specificity in enzymes important for biotechnology or industrial processes and the enzymes can then be engineered to accept other substrates of interest (for example see Blikstad *et al.*, 2014). Moreover, molecular modeling can be a key method for understanding and engineering the biophysical properties of enzymes and also for predicting protein-protein interactions through computational docking (Schwede *et al.*, 2009). The determining factor for how a protein model can be used is ultimately the accuracy and quality of the model. Drug design requires a model of high accuracy and reliability, especially in the ligand-binding site, to be able to discern valuable information about protein-ligand interactions (Kopp *et al.*, 2007; Thorsteinsdottir *et al.*, 2006). A good accuracy is also attractive when studying the effects of mutations, where a specific site is in focus and differences in side chain properties might have significant implications on the function and folding of the protein. These types of analyses can explain experimental results or guide point mutation experiments.

2.3.1 Docking

Today, docking is an essential tool both for the development of new pharmaceuticals and for studying protein-protein interactions (Chen, 2015). The biggest use of docking tools is exploited for studying how a small molecule binds to a protein, particularly in computer-aided drug design (Merz *et al.*, 2010; Pei *et al.*, 2014; Warren *et al.*, 2012; Zheng *et al.*, 2013). The freely available Autodock (Goodsell & Olson, 1990) and the commercial GOLD (Jones *et al.*, 1995; Jones *et al.*, 1997) and Glide (Schrödinger, LLC) (Friesner *et al.*, 2004; Halgren *et al.*, 2004), among others, are examples of popular docking programs (Azam & Abbasi, 2013; Ferreira *et al.*, 2015; Lape *et al.*, 2010; Li *et al.*, 2014). Molecular docking is a method used to predict protein-ligand interactions by sampling through many docked conformations of the ligand to find the one specific conformation, which corresponds to the right/real one. This conformation is the global minimum of the energy landscape and can be identified through exploring the various potential binding modes of the ligand and predicting the interaction energy of each of these modes (Kapetanovic, 2008). Sampling through the different ligand conformations includes changing the structural parameters, such as torsional, translational and rotational bonds (Ferreira *et al.*, 2015; Guedes *et al.*, 2014). This can be achieved by implementing either systematic or stochastic search methods. Systematic algorithms explore all degrees of freedom in the ligand, after which it converges to the most likely binding mode, *i.e.* the minimum energy conformation. The obvious problem with this algorithm is that the

more the rotational freedom for the ligand increases, the more possible combinations have to be taken into account. To help with this, the incremental construction algorithm makes it possible to break the ligand structure into several fragments and then build it back up in the binding site (Ferreira *et al.*, 2015). Here, only one fragment at a time is considered and, therefore, reduces the degrees of freedom, which have to be accounted for. It starts from one anchor fragment, which is first docked into the binding site and the other fragments are then sequentially added until the ligand is complete. On the other hand, stochastic methods randomly change the structural parameters of the ligand and generate a diverse set of solutions. Each pose is then evaluated and either rejected or not (Ferreira *et al.*, 2015; Guedes *et al.*, 2014). The program GOLD, which has been used in this work, uses an application of the stochastic methods called genetic algorithms (Jones *et al.*, 1995; Jones *et al.*, 1997). This method encodes all the structural parameters in a chromosome and starting from this, the random search algorithm generates a population (Ferreira *et al.*, 2015). The energy values of the population members are evaluated and the ones with lowest energy serve as templates for the generation of the next population and the procedure is repeated until it converges to the global energy minimum conformation.

Docking approaches

The original thought behind docking was to fit a key (the ligand) into a lock (the protein) and including protein flexibility in the search for an optimal complex is still a challenge (Guedes *et al.*, 2014). However, progress has been made in this aspect and the methods used to account for protein flexibility are mainly divided into five classes: soft docking, side-chain flexibility, molecular relaxation, ensemble docking and collective degrees of freedom (Teodoro & Kavraki, 2003). These methods consider that sometimes the shape of the binding pockets will change upon ligand-binding and in these cases rigid docking is not sufficient enough to give valuable answers (Chang *et al.*, 2011; Chen *et al.*, 2014; Tou *et al.*, 2013; Tou & Chen, 2014). However, this factor is taken into account at the expense of time and computational power. The soft docking approach addresses small conformational changes in the protein by allowing small overlaps between the protein and the ligand atoms (Jiang & Kim, 1991). When the induced fit effect is bigger than can be handled by soft docking, the side-chain flexibility approach can be used (Leach, 1994). This keeps the protein backbone fixed but the side-chains are allowed to change conformation, which in turn gives a favorable and tight binding (Koshland *et al.*, 1966). Molecular relaxation, on the other hand, is usually a post-processing step after docking a ligand to a rigid protein to relax the complex and allow the protein backbone and side-chains, as well as the ligand to move and adjust to each other (Armen *et al.*, 2009; Nowosielski *et al.*, 2013). Thus, this method is mostly used as a refinement step for the docking poses and to evaluate the stability of the

complex. Alternatively, the protein conformations can be considered as an ensemble; similar to the way they behave in solution (Nichols *et al.*, 2011; Novoa *et al.*, 2010; Sperandio *et al.*, 2010). One of these conformations should match the shape of the ligand and, therefore, the ligand can bind and shift the equilibrium of the protein conformations to the ligand bound state (Kumar *et al.*, 2000; Ma *et al.*, 2002; Tsai *et al.*, 1999a; Tsai *et al.*, 1999b). Additionally, there is a possibility to work with full protein flexibility through the collective degrees of freedom method (Teodoro *et al.*, 2003). This method tries to capture the dominant protein motion forms and dock the ligand to these. The choice between which of these methods to use depends on the accessible computing power and the target protein and its behavior. If the aim is to screen a database with millions of drug molecules against a target protein, then rigid docking is the smartest choice with regard to time. But if there are only a few compounds, which should be docked to one pocket on the protein, several restrictions are imposed and flexible docking might be reasonable to optimize the analysis of the interactions (Chen, 2015). Furthermore, the flexible docking approach benefits from knowledge or experimental data on residues involved in ligand-binding, since it allows for restriction of the flexibility to only these specific residues.

Scoring functions

Docking results give many different options for protein-ligand or protein-protein interaction modes. Different types of scoring functions evaluate the docking poses and enables ranking of the results. Scoring functions have traditionally been divided into force-field-based methods, empirical scoring functions and knowledge-based functions (Böhm & Stahl, 2002; Muegge & Rarey, 2001). However, Liu & Wang (2015) highlight that these terms are not up to date and their varied use in the literature can cause confusion. Therefore, Liu & Wang have made an effort to produce a naming convention for the scoring functions and suggest the following names and groups: physics-based methods, empirical scoring functions, scoring functions described as knowledge-based potential and descriptor-based scoring functions (Liu & Wang, 2015). The first group, physics-based methods, is based on force fields to compute van der Waals and electrostatic energies, *i.e.* noncovalent interactions, between the protein and the ligand. They are often improved by adding other methods to take into account solvation energies. GoldScore implemented in the GOLD docking program (Jones *et al.*, 1995; Jones *et al.*, 1997) is an example of a physics-based method. Empirical scoring functions consider multiple individual terms, which are important in protein-ligand binding (Liu & Wang, 2015). It has rewarding scores for favorable interactions, such as hydrogen bonds, coordinate bonds and lipophilic contacts, while it penalizes for example steric clashes and frozen rotatable bonds. Protein-ligand complexes with experimentally solved 3D structures and known data for binding affinity are then used for regression

analysis to derive a weight factor for each individual term, after which they are summed together to compute the fitness of the protein-ligand binding. Examples of empirical scoring functions are PLP (Gehlhaar *et al.*, 1995; Verkhivker *et al.*, 1995), ChemScore (Eldridge *et al.*, 1997; Murray *et al.*, 1998) and GlideScore-XP (Friesner *et al.*, 2006), which is considered as the most sophisticated emipirical scoring function at present. Scoring functions described as knowledge-based potential take into account pairwise contacts between the protein and the ligand (Liu & Wang, 2015). They are derived through statistical analysis of structural information from known 3D structures of protein-ligand complexes, but do not consider experimental binding data. The frequency of a pairwise contact is considered as its energetic contribution to the protein-ligand complex, which means that an energetically favorable interaction between a pair of atoms occurs often. On the other hand, if a certain pairwise contact is rarely seen, it indicates a less favorable interaction. An example of such a scoring function is DrugScore (Gohlke *et al.*, 2000; Neudert & Klebe, 2011; Velec *et al.*, 2005). The group consisting of descriptor-based scoring functions is a new trend and introduces quantitative structure-activity relationship analysis into the evaluation of protein-ligand interactions (Liu & Wang, 2015). These types of scoring functions encode the protein-ligand interaction patterns in descriptors, and machine-learning techniques are then applied to obtain statistical models, which can compute protein-ligand scores. One example of a descriptor-based scoring function is NNScore (Durrant & McCammon, 2010; Durrant & McCammon, 2011). The scoring functions are evaluated in benchmarks, CASF (Cheng *et al.*, 2009; Li *et al.*, 2014a; Li *et al.*, 2014b) and CSAR (Damm-Ganamet *et al.*, 2013; Dunbar *et al.*, 2013; Smith *et al.*, 2011) and, currently, it is suggested that the prediction of the correct ligand pose is well handled by the scoring functions. However, the prediction of protein-ligand binding affinities and, hence, the ranking of the ligands still requires attention. So far, ChemPLP (Korb *et al.*, 2009) in GOLD (Jones *et al.*, 1995; Jones *et al.*, 1997) and PLP (Gehlhaar *et al.*, 1995; Verkhivker *et al.*, 1995) in Discovery Studio (Accelrys Inc.) have both shown a good balance between docking and ranking power (Li *et al.*, 2014a).

2.4 Modeling success stories

2.4.1 Leptin and its receptor

One of the key players for regulation of body weight is the leptin protein and its receptor, which cause a signal affecting the food intake and energy expenditure (Ingalls *et al.*, 1950). The obese gene codes for leptin (Zhang *et al.*, 1994) and at the time it was sequenced it did not show significant sequence similarity to any other known protein. To study the mechanism of action, Bryant and colleagues used a fold recognition method and suggested a structure similar to cytokines, comprising a bundle of four α-helices (Madej

et al., 1995). This led to a prediction that leptin, similarly to cytokines, exerts its signaling through a receptor, which could possibly be mutated in obese humans. In 1995, the leptin receptor was found (Tartaglia *et al.*, 1995) and in 1997, the crystal structure of human leptin (PDB code 1AX8) (Zhang *et al.*, 1997) confirmed the model and the similarity to cytokines. It would last until 2012 before the crystal structure of the leptin-binding domain of the leptin receptor was solved (PDB code 3V6O) (Carpenter *et al.*, 2012) but, before this, the same domain was modeled (Iserentant *et al.*, 2005; Niv-Spector *et al.*, 2005a; Niv-Spector *et al.*, 2005b) and used for docking studies with leptin, which has given reasonable structural explanations for obtained experimental results (Peelman *et al.*, 2014). Hence, structural modeling has provided the leptin research with valuable information and helped the researchers find key data. The next aim is to study the activation of the leptin receptor with the help of computational and experimental methods (Mancour *et al.*, 2012; Moharana *et al.*, 2014) in order to explain obesity and develop drugs working through the leptin receptor (Peelman *et al.*, 2014; Tramontano, 2006).

2.4.2 G protein-coupled receptors

G protein-coupled receptors (GPCRs) are a large family of signal transducing proteins, which makes them very interesting drug targets and pharmaceutical research on these proteins is intense (Carlsson *et al.*, 2011). The number of drugs targeting GPCRs reflects the importance of these proteins – almost 30 % of all approved drugs are dedicated to work through GPCRs (Overington *et al.*, 2006). For a long time, bovine rhodopsin was the only available crystal structure from this protein family (PDB code 1F88) (Palczewski *et al.*, 2000) and it was frequently used to model the structure of other family members, which were then used for docking studies (Bissantz *et al.*, 2005; de Graaf *et al.*, 2008; Engel *et al.*, 2008; Kellenberger *et al.*, 2007; Kratochwil *et al.*, 2005; Kurczab *et al.*, 2010; Michino *et al.*, 2009; Salo *et al.*, 2005; Shi & Javitch, 2002; Tikhonova *et al.*, 2008). The lack of sequence identity poses challenges (Li *et al.,* 2010) for these types of studies, but the structures are highly similar with seven transmembrane α-helices coupled together by loops. Later on, the crystal structures of other family members have been solved, thereby providing an opportunity to compare and verify the structural models (Hanson & Stevens, 2009). This comparison has proven the modeling and docking studies to be effective, and still today homology modeling of GPCRs is essential due to a significant lack of experimentally solved crystal structures (Carlsson *et al.*, 2011).

2.4.3 The HIV protease

In 1981, the U.S. Centers for Disease Control reported the first cases of an infection with subsequent collapse of the immune system and the illness was

defined as acquired immunodeficiency syndrome (AIDS). The researchers faced a global pandemic and major efforts were made in the 1980s to find a cure for the disease. Within three years of the reported cases, a single-stranded RNA virus from the *Lentiviridae* family was found responsible for causing the disease and it is now called human immunodeficiency virus (HIV) (Barre-Sinoussi *et al.*, 1983; Coffin *et al.*, 1986; Gallo *et al.*, 1984; Gallo & Montagnier, 1988; Popovic *et al.*, 1984). The genome of the HIV virus was sequenced two years after it was found to be the causing agent of AIDS (Ratner *et al.*, 1985). This revealed an Asp-Thr(Ser)-Gly triad, which can also be found in the aspartic acid protease family (Power *et al.*, 1986; Toh *et al.*, 1985). The proteins of this family are built from two homologous domains, with each domain contributing its own catalytic triad to the active site (Pearl & Blundell, 1984; Tang *et al.*, 1978). Moreover, the two Asp residues coordinate a water molecule believed to be important for the activity (Pearl, 1987). However, Pearl and Taylor (1987) found that the HIV protease did not contain two catalytic triads and only half the number of expected amino acids, hence, suggesting that this protease was a single domain protein (Pearl & Taylor, 1987). With the help of structural modeling of the HIV protease based on the crystal structure of endothiapepsin (PDB code 4APE) (Pearl & Blundell, 1984) they found that it is possible for the HIV protease to exhibit the aspartic acid protease fold. Hence, they deduced that the single domain HIV protease represents an ancestral protease, from which the aspartic acid proteases with two domains have evolved and, therefore, the HIV protease might form a dimer to be able to make all the important interactions needed for an aspartic acid protease. As a consequence, they modeled the dimeric structure of the HIV protease and showed that it provides the right active site structure and a distinct substrate binding cleft (Pearl & Taylor, 1987). Thereafter, this has been confirmed by a number of methods, including X-ray crystallography (Lapatto *et al.*, 1989; Navia *et al.*, 1989; Wlodawer *et al.*, 1989). Based on all the structural data in combination with experimental strategies, the researchers were able to design potent treatment for the disease within a decade of its finding, thereby making the disease chronic rather than fatal (Huff & Kahn, 2001).

3 Aims of the study

The aim of this thesis was to create a workflow for guidance through the process of protein structural modeling, with emphasis on critical steps and important facts to take into account to avoid errors in the final structural model. Through separate case studies with different focus, the possibilities and importance of structural bioinformatics in protein structure and function research are highlighted. The publications contributing to the thesis will be discussed individually.

In publication I and II, the aim was to create 3D models for the *Escherichia coli* propanediol oxidoreductase (FucO) mutants and dock experimentally tested ligands to them. The purpose was to give structural explanations for the experimentally observed differences in substrate scope, with the ultimate goal of finding an enzyme variant with efficient catalysis of *S*-3-phenyl-1,2-propanediol, which could be used in the pharmaceutical industry for the production of drug components.

In publication III and IV, the aim was to create 3D models for *Yersinia enterocolitica* LpxR and *Klebsiella pneumoniae* LpxO and find amino acids, which are important for their catalytic activities and substrate specificities. Together with the experimental data, this would give valuable information about the disease-causing mechanisms in these bacteria and ultimately guide drug development.

In publication V, the aim was to model the 3D structure of the *Synechocystis* PCC 6803 Slr0006 protein in order to identify a putative active site and amino acids, which would play a key role for the protein function. This protein is inadequately characterized both structurally and functionally, but it is, however, very interesting from a climate point of view to show how cyanobacteria and higher plant organisms can adapt to new conditions in the environment.

In publication VI, the aim was to model the 3D structure of human cancerous inhibitor of protein phosphatase 2A (CIP2A) to get a structural insight into how this protein and its cancer-causing mechanism could be inhibited with therapeutics. So far, it has been established that CIP2A is involved in different cancer types, which makes it an important drug target, but the lack of structural data greatly hampers the design of new therapeutics targeting this protein. Hence, all structural data would greatly benefit the attempts of finding an anti-cancer drug asserting its effects through CIP2A.

4 Methods

4.1 Sequence and structural data

The amino acid sequences for the proteins to be modeled were obtained from UniProtKB (UniProt Consortium, 2015) and then used as baits in searches with BLAST (Altschul *et al.*, 1990) at NCBI. With the standard protein-protein BLAST program (blastp), UniProtKB and the non-redundant sequence database (GenBank, Refseq, PDB, SwissProt, PIR, PRF) were searched to obtain homologous sequences, while PDB (Berman *et al.*, 2000; Berman *et al.*, 2002) was used as search database for obtaining crystal structures suitable as templates for homology modeling.

4.2 Sequence analysis

The secondary structure profiles for the target proteins were predicted with the APSSP2 (publication VI) (Raghava, 2002), PSIPred (publication VI) (Jones, 1999), PORTER (publication VI) (Mirabello & Pollastri, 2013; Pollastri & McLysaght, 2005) and JPred (publication V and VI) (Cole *et al.*, 2008) servers. The CIP2A amino acid sequence was also analyzed with SMART (publication VI) (Letunic *et al.*, 2015; Schultz *et al.*, 1998) to determine the different domains in the protein. Furthermore, the amino acid sequence for KpLpxO was analyzed with the transmembrane helix prediction server TMHMM (publication IV) (Sonnhammer *et al.*, 1998) to determine the soluble, *i.e.* not transmembrane, portions of the protein, which consequently are possible to model based on a related, crystallized protein.

4.3 Sequence alignment

Global sequence alignments were performed with the programs MALIGN (Johnson *et al.*, 1996) and VERTAA (Johnson & Lehtonen, 2000) in the BODIL modeling environment (Lehtonen *et al.*, 2004). MALIGN was used for both pairwise (publication III) and multiple sequence alignment (publications IV, V, VI) with STRMAT110 as scoring matrix and a gap penalty of 40. In publication IV and VI, multiple sequence alignments were generated between the protein of interest and homologous sequences and the alignments were manually inspected and edited. In publication IV, the amino acid sequence for the human Asp/Asn β-hydroxylase used as template was aligned to the prealigned multiple sequence alignment and all sequences except LpxO and human Asp/Asn β-hydroxylase were deleted from the alignment before modeling. In publication V, VERTAA (Johnson & Lehtonen, 2000) was used to superimpose the structures of proteins from the Sua5/YrdC/YciO family and generate a structure-based multiple sequence alignment from the superimpositions. The Slr0006 sequence was then aligned

with MALIGN (Johnson *et al.*, 1996) to the prealigned structure-based alignment. For modeling, all sequences except Slr0006 and the amino acid sequence for the structure to be used as template were deleted.

4.4 Modeling of 3D structure

The homology models of the target proteins FucO, LpxR, LpxO and Slr0006 (publications I – V) were produced with MODELLER (Sali & Blundell, 1993), which uses the satisfaction of spatial restraints method for model building. Ten models were generated for each target protein, and the models with the lowest energy according to the MODELLER objective function were chosen for further studies.

In publication VI, no homologous proteins with known structure were found and, therefore, we approached the modeling of the CIP2A armadillo domain through the available modeling servers I-TASSER (Roy *et al.*, 2010; Yang *et al.*, 2015; Zhang, 2008b), Phyre (version 2.0) (Kelley & Sternberg, 2009) and HHPred (Remmert *et al.*, 2011; Söding, 2005; Söding *et al.*, 2005). I-TASSER (Roy *et al.*, 2010; Yang *et al.*, 2015; Zhang, 2008b) uses LOMETS (Wu & Zhang, 2007), which is a multiple-threading program, to identify structural templates from PDB (Berman *et al.*, 2000) and constructs models based on iterative template fragment assembly simulations. The models are then threaded through the protein function database BioLiP (Yang *et al.*, 2013) to obtain functional information. Phyre (Kelley & Sternberg, 2009) and HHPred (Remmert *et al.*, 2011; Söding, 2005; Söding *et al.*, 2005) use the principles of homology modeling and start by detecting distant homologs to be used as templates for the modeling, if the sequence alignment shows a relationship between the target and the template.

4.5 Model analysis

The programs and servers PROCHECK (publication V, VI) (Laskowski *et al.*, 1993), WHATCHECK (publication V) (Hooft *et al.*, 1996), ProSA-web (publication IV, V, VI) (Sippl, 1993; Wiederstein & Sippl, 2007), QMEAN (publication IV, VI) (Benkert *et al.*, 2009), ERRAT (publication VI) (Colovos & Yeates, 1993), Verify-3D (publication VI) (Bowie *et al.*, 1991; Eisenberg *et al.*, 1997; Luthy *et al.*, 1992), ProQ (publication VI) (Wallner & Elofsson, 2003) and MODFOLD4 (publication IV, VI) (Buenavista *et al.*, 2012; McGuffin & Roche, 2010; McGuffin *et al.*, 2013) were used to assess the quality of the models. Furthermore, all models were superimposed on the known crystal structures of the templates and visually examined and evaluated. The conformation of the Leu225 – Asn230 loop in the model of the CIP2A armadillo domain (CIP2A-ArmRP; publication VI) was optimized using Loopy in JACKAL (Petrey *et al.*, 2003; Xiang & Honig, 2001; Xiang *et al.*, 2002), while the conformation of Asp31 in LpxR was optimized with

JACKAL rotamer search (publication III). In both cases, the conformation with the lowest energy was chosen as the final conformation.

In publications V and VI, the distribution of electrostatic charges on the surface of the Slr0006 and CIP2A-ArmRP models was calculated with the APBS tool in PyMOL (version 1.4, Schrödinger, LLC). Furthermore, in publications III and IV, SURFNET (Laskowski, 1995) was used to calculate cavities and amino acids lining the cavities in LpxR and LpxO. For publication VI, possible ligand-binding cavities in CIP2A were searched with MetaPocket 2.0 (Huang, 2009; Zhang *et al.*, 2011) and ConSurf (Ashkenazy *et al.*, 2010; Celniker, 2013; Glaser *et al.*, 2003; Landau *et al.*, 2005) was used to generate a multiple sequence alignment as basis for mapping conserved areas onto the surface of the modeled CIP2A-ArmRP.

4.6 Molecular docking

In publications I and II, the structure for the phenylacetaldehyde ligand was taken directly from the crystal structure of *E. coli* amine oxidase (PDB code 1D6U) (Wilmot *et al.*, 1999), while *S*-3-phenyl-1,2-propanediol was derived by editing (2*R*, 3*S*)-3-amino-3-phenylpropane-1,2-diol from the crystal structure of *Scytalidium lignicola* Scytalidopepsin B (PDB code 2IFR) (Pillai *et al.*, 2007) with Maestro Molecular Modeling Interface (Version 9.3., Schrödinger, Inc.). The crystal structure of wild type FucO and the models of the mutants were prepared for docking in Discovery Studio, along with the ligands. In the active site of the protein, O7N was constrained as an acceptor of hydrogen bonds, while H23 and H25 of NAD^+ were defined as hydrogen bond donor and acceptor, respectively. In phenylacetaldehyde, O9 was a possible acceptor of hydrogen bonds, while in *S*-3-phenyl-1,2-propanediol O1 and O5 were constrained as acceptors, and H22 and H23 as hydrogen bond donors. GOLD via Discovery Studio was used to dock the ligands to the rigid protein receptors. The docking poses were analyzed and scored with the Score Ligands function in Discovery Studio and the poses with the highest PLP2 score were chosen as the best poses.

For the docking studies in publication III, the Kdo_2-lipid A ligand was modified from the coordinates of the lipopolysaccharide (LPS) molecule bound to FhuA (PDB code 2FCP) (Ferguson *et al.*, 1998) with the program SYBYL (version 8.0, Tripos Associates, Inc., St Louis, MO, USA). The fatty acyl chains were removed to make the docking easier with fewer rotatable bonds and aminoarabinose was added to the structure, after which it was minimized with the conjugate gradient method and Tripos force field. The modified Kdo_2-lipid A ligand with and without aminoarabinose was docked to the YeLpxR structural model and the crystal structure of StLpxR (PDB code 3FID) (Rutten *et al.*, 2009) with GOLD via Discovery Studio (Version 3.5., Accelrys Inc.). Default parameters were used and the receptor cavity

was defined to amino acids Asp10, Gln16, Thr/Ser34, Lys67, and Tyr130. The receptor conformations were kept rigid throughout the docking.

4.7 Molecular dynamics simulations

In publication VI, the final model was validated to have a stable fold with parallel molecular dynamics (MD) simulations. These simulations allow for analyses of how atoms in the protein behave and change over time. Energy minimization, thermal equilibration and standard production simulations for the CIP2A-ArmRP model were done with the AMBER package (v.12) (Case *et al.*, 2012) and the AMBER ff03 force field (Duan *et al.*, 2003). All simulations were run in an octahedral box filled with explicit TIP3P water molecules (Jorgensen *et al.*, 1983) and extending 10 Å from the protein. Six neutralizing Na^+ ions were added for the model, while the template structure required 15 Na^+ ions for neutralization. Periodic boundary conditions and particle-mesh Ewald electrostatics (Essmann *et al.*, 1995) were used, while the cut-off for non-bonded interactions was 9 Å. For Langevin dynamics during simulation, a 1 fs or 2 fs time step was applied and the hydrogen atoms were constrained with the SHAKE algorithm (Ryckaert *et al.*, 1977). The temperature and pressure were held constant at 300 K (coupling constant 5.0 ps) and 1 bar (coupling constant 2.0 ps), respectively, during the 20 ns production simulations (Berendsen *et al.*, 1984). VMD (Humphrey *et al.*, 1996) and the ptraj module in AMBER were used to study the trajectories, while PyMOL was used to study the final model.

4.7.1 Energy minimization

Steepest descent and conjugate gradient methods were used in six steps and the restraints on the atoms to their initial position was gradually reduced during these steps. A maximum of 200 iterations was defined for each step, of which the ten first iterations were performed with the steepest descent method and then the conjugate gradient method was applied. The restraint force constant at each step was 10, 5, 1, 0.1, 0.01 and 0 kcal/mol$Å^2$.

4.7.2 Equilibration simulations

Five steps were used for the equilibrium simulations:
1) A Langevin thermostat, collision frequency (γ) of 1.0 ps^{-1}, constant volume and a force constant of 5 kcal/mol$Å^2$ to restrain the protein atom positions were used for heating the system from 10 K to 300 K for 10 ps.
2) Same as previous step but without restricting the protein atom positions and for 20 ps instead of 10 ps.
3) 20 ps MD at 300 K, Langevin thermostat, $\gamma = 0.5$ ps^{-1}, constant volume, no restraints on the protein atoms.
4) 50 ps MD at 300 K, Langevin thermostat, $\gamma = 0.5$ ps^{-1}, constant pressure of 1.0 bar, pressure coupling constant 1.0 ps, no restraints on the protein atoms.

5) 400 ps MD at 300 K, constant pressure of 1.0 bar, pressure coupling constant 2.0 ps, temperature coupling constant 5.0 ps, no restraints on the protein atoms.

4.8 Visualization

High-resolution pictures of the structural models in each publication were made with PyMOL (version 1.4, Schrödinger, LLC) and labels were added in the GNU Image Manipulation Program (version 2.6.9). Pictures of the sequence alignments in publications III, IV, V and VI were made with ESPript 2.2 (Gouet *et al.*, 1999).

4.9 Experimental work

Our collaborators performed all the experimental work contributing to this thesis in their laboratories. The experiments are explained in detail in the Materials and Methods section of the original publications I, II, III, IV and V.

Table 1 Web Resources

RESOURCE	URL	REFERENCE
	SEQUENCE AND STRUCTURAL DATA	
UniProtKB	http://www.uniprot.org/	UniProt Consortium, 2015
BLAST	http://blast.ncbi.nlm.nih.gov/	Altschul *et al.,* 1990
PDB	http://www.pdb.org/	Berman *et al.,* 2000
	SEQUENCE ANALYSIS	
APSSP2	http://www.imtech.res.in/raghava/apssp2/	Raghava, 2002
PSIPred	http://bioinf.cs.ucl.ac.uk/psipred/	Jones, 1999
PORTER	http://distill.ucd.ie/porter/	Mirabello & Pollastri, 2013; Pollastri & McLysaght, 2005
JPred	http://www.compbio.dundee.ac.uk/www-jpred/	Cole *et al.,* 2008
SMART	http://smart.embl-heidelberg.de/	Letunic *et al.,* 2015; Schultz *et al.,* 1998
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Sonnhammer *et al.,* 1998
	THREE-DIMENSIONAL STRUCTURAL MODELING	
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Roy *et al.,* 2010; Yang *et al.,* 2015; Zhang, 2008
LOMETS	http://zhanglab.ccmb.med.umich.edu/LOMETS/	Wu & Zhang, 2007

Table 1 Continued

RESOURCE	URL	REFERENCE
Phyre	http://www.sbg.bio.ic.ac.uk/phyre2	Kelley & Sternberg, 2009
HHPred	http://toolkit.tuebingen.mpg.de/hhpred/	Remmert *et al.,* 2011; Söding, 2005; Söding *et al.,* 2005
BioLip	http://zhanglab.ccmb.med.umich.edu/BioLip/	Yang *et al.,* 2013
	MODEL ANALYSIS	
ProSA-web	http://prosa.services.came.sbg.ac.at/prosa.php	Sippl, 1993; Wiederstein & Sippl, 2007
QMEAN	http://swissmodel.expasy.org/qmean/	Benkert *et al.,* 2009
ERRAT	http://services.mbi.ucla.edu/ERRAT/	Colovos & Yeates, 1993
Verify-3D	http://services.mbi.ucla.edu/Verify_3D/	Bowie *et al.,* 1991; Eisenberg *et al.,* 1997; Luthy *et al.,* 1992
ProQ	http://www.sbc.su.se/~bjornw/ProQ/	Wallner & Elofsson, 2003
MODFOLD4	http://www.reading.ac.uk/bioinf/ModFOLD_form_4_0.html	Buenavista *et al.,* 2012; McGuffin & Roche, 2010; McGuffin *et al.,* 2013
JACKAL	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Jackal	Petrey *et al.,* 2003; Xiang & Honig, 2001; Xiang *et al.,* 2002
MetaPocket 2.0	http://projects.biotec.tu-dresden.de/metapocket/	Huang, 2009; Zhang *et al.,* 2011
ConSurf	http://consurf.tau.ac.il	Ashkenazy *et al.,* 2010; Celniker, 2013; Glaser *et al.,* 2003; Landau *et al.,* 2005

5 Results and discussion

5.1 FucO (*E. coli*)

5.1.1 Introduction

In the cell, NAD(P)H-dependent dehydrogenases stereoselectively catalyze the oxidation of primary and secondary alcohols into α-hydroxyaldehydes and ketones, which are important building blocks for natural products and synthetic drugs (Adams & Levine, 1923; Enders & Bhushan, 1988; Hoyos *et al.*, 2010; Kratzer & Nidetzky, 2007). Industrial synthesis of these molecules requires harsh conditions and, therefore, the power and stereoselectivity of the natural enzymes producing the chiral molecules have turned them into targets for use in biocatalysis (Blank *et al.*, 2010; Goldberg *et al.*, 2007; Hall & Bommarius, 2011; Monti *et al.*, 2011). Naturally, enzymes have specificity towards a certain substrate, but a broader substrate scope can be achieved by re-engineering and, hence, make the enzymes better biocatalysts (Bornscheuer *et al.*, 2012). Our enzyme of interest, FucO, is a homodimeric class III Fe^{2+} dependent alcohol dehydrogenase (Montella *et al.*, 2005; Reid & Fewson, 1994). Each FucO subunit is 41 kDa with an all α-helical C-terminal domain and an N-terminal domain, which has a α/β-dinucleotide binding fold (PDB code 2BL4) (Montella *et al.*, 2005). A tunnel in between the two domains defines the active site, where a Fe^{2+} ion is coordinated by three histidines (His200, His263, His277) and one aspartate (Asp196). FucO converts *S*-lactaldehyde to *S*-1,2-propanediol (Figure 4a), and *vice versa*, in the catabolic pathway of fucose and rhamnose (Baldoma & Aguilar, 1988; Boronat & Aguilar, 1979; Conway & Ingram, 1989). It is highly specific for aliphatic, low molecular weight primary 2-*S* alcohols, it can be easily produced and the tertiary structure is known, which makes it an ideal target for re-engineering towards a biocatalyst (Blikstad & Widersten, 2010). In publication I and II, we set out to create FucO mutants, which would catalyze the conversion of *S*-3-phenyl-1,2-propanediol (Figure 4b) into the corresponding α-hydroxyaldehyde, which is an important intermediate for industrial synthesis of pharmaceuticals, fine chemicals and natural products. *S*-3-phenyl-1,2-propanediol is bigger and bulkier than the natural substrate due to an additional phenyl ring and, therefore, our objective was to enlarge the active site by analysis and targeted mutation of amino acids known to be catalytically important.

5.1.2 Target substrate vs. screening substrate

The available crystal structure of FucO (PDB code 2BL4) (Montella *et al.*, 2005) enabled determination of the amino acids, which restrict the active site cavity. These amino acids were then the targets for iterative saturation mutagenesis to create mutants with a larger active site. Phenylacetaldehyde

(Figure 4c) was included as a surrogate substrate in the activity screening, since the reduction reaction is faster than the oxidation reaction (Blikstad & Widersten, 2010) and, therefore, could facilitate the detection of even weakly active variants. Additionally, the product from *S*-3-phenyl-1,2-propanediol oxidation is not commercially available and, therefore, a structural analog is a logical choice for performing a reduction reaction. Phenylacetaldehyde lacks the *sec*-alcohol at the α-position and a methylene group compared to oxidized *S*-3-phenyl-1,2-propanediol but, despite this, phenylacetaldehyde was deemed to be a reasonably good analog. Nevertheless, kinetic characterization showed that the best catalyst of phenylacetaldehyde, the Asn151Gly-Leu259Val mutant with a 4400-fold increase in k_{cat}/K_M compared to wild type FucO, had lost activity with the target substrate *S*-3-phenyl-1,2-propanediol. The opposite was true for the best catalyst of *S*-3-phenyl-1,2-propanediol. This Val164Cys-Leu259Val-Cys362Gly mutant had a 43-fold increase in turnover of the target substrate compared to wild type FucO, but it did not show activity with the screening substrate phenylacetaldehyde. Hence, in publication I, each of these mutants was modeled and used for docking studies with the respective substrate to find structural reasons for the substrate specificity.

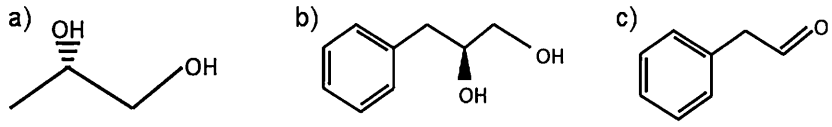


Figure 4. FucO substrates. a) The structure of the natural substrate *S*-1,2-propanediol. b) The structure of the target substrate *S*-3-phenyl-1,2-propanediol. c) The structure of the screening substrate phenylacetaldehyde.

5.1.3 Asn151 and Phe254 are key residues for substrate specificity

The modeling and docking results show that both the Asn151Gly-Leu259Val mutant and the Val164Cys-Leu259Val-Cys362Gly mutant are able to accommodate the bulkier phenyl-containing screening and target compounds, respectively, which means that the increase in active site size enables catalysis (Figure 5). The aldehyde oxygen of phenylacetaldehyde hydrogen bonds to the amide of the NAD^+ cofactor in the Asn151Gly-Leu259Val mutant and the bent conformation allows the phenyl ring to π-π stack with Phe254 (Figure 5a). The latter of these interactions is lost in the wild type protein due to Asn151 (Figure 5b). The side chain of this amino acid points straight into the active site, thereby restricting the active site volume together with Leu259 on the opposite side, which makes it impossible for the phenyl ring to be accommodated in the active site. Hydrogen bonds are formed

between the aldehyde oxygen and the cofactor, leaving the phenyl ring to protrude from the active site cavity and the complex is not stable. Therefore, wild type FucO is unable to efficiently catalyze phenylacetaldehyde. The target substrate *S*-3-phenyl-1,2-propanediol also binds in a bent conformation to the Val164Cys-Leu259Val-Cys362Gly mutant and forms a π-π stacking interaction between the phenyl ring of the substrate and Phe254 (Figure 5c). However, Asn151 is retained in this mutant, but instead the π-π stacking interaction is enabled by the Leu259Val mutation, which creates the required additional space for the phenyl ring to be accommodated in the active site. The other mutations, Val164Cys and Cys362Gly, enable both hydroxyl groups in the *S*-3-phenyl-1,2-propanediol molecule to form hydrogen bonds to the cofactor, while one of them also creates a hydrogen bond to Asn151, which makes this an important amino acid for the formation of a stable complex for catalysis. Consequently, for the mutants to be active with phenylacetaldehyde, they require mutation of Asn151, while the same residue needs to be retained for the formation of a stable complex between the mutants and *S*-3-phenyl-1,2-propanediol. Furthermore, docking of the target substrate to wild type FucO shows that the lack of space for a phenyl-substituted compound is the main reason for the lack of activity with the target substrate (Figure 5d).

Although it has been shown in other studies that coevolution of highly specialized proteins can be successful by iteratively selecting for activity towards structural intermediates of the ultimate target ligand (Chen & Zhao, 2005) the results show that phenylacetaldehyde was not the best screening substrate considering that the mutants, which showed activity with this substrate did not have activity with *S*-3-phenyl-1,2-propanediol. The responsible factor, Asn151, is crucial for the hydrogen-bonding network of the target substrate to form a stable complex for catalysis, while the same residue inflicts steric hindrance for the accommodation and efficient catalysis of the screening substrate phenylacetaldehyde. Additionally, the docking results highlight Phe254 as an important residue for the formation of stable complexes with both substrates, and this residue is retained in all mutants displaying improved catalytic activity compared to wild type FucO. The structural analysis implies an important role for its aromatic side chain and its ability to form π-π stacking interactions.

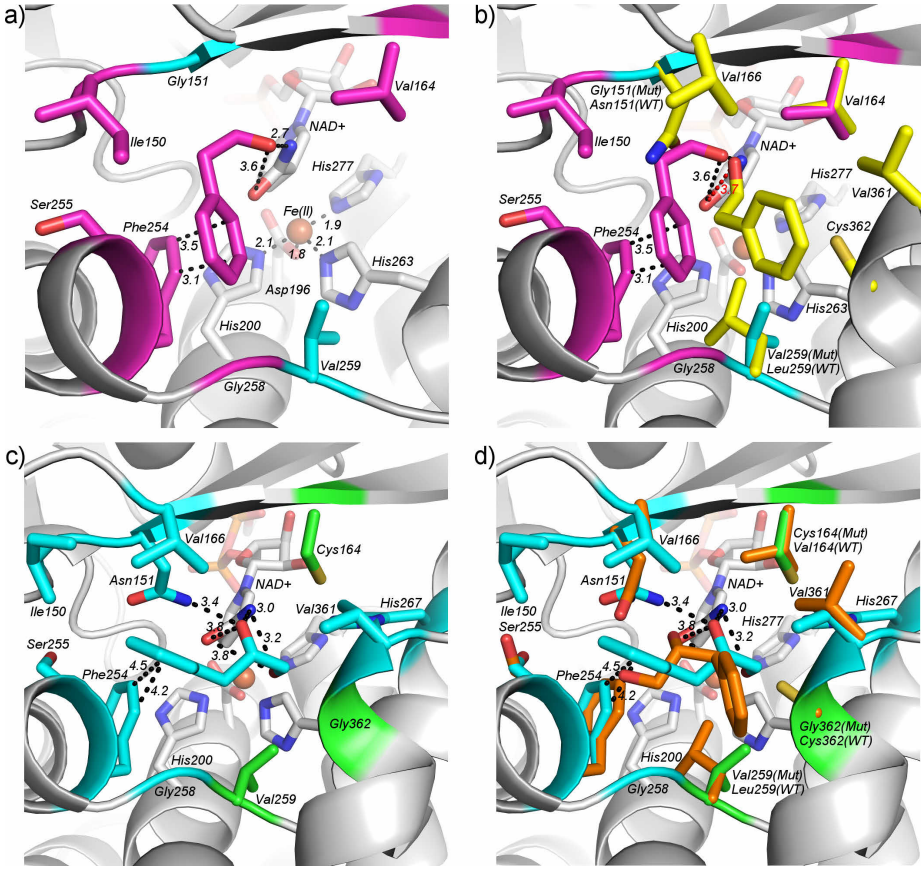


Figure 5. The most active FucO mutants. The Asn151Gly-Leu259Val mutant has the highest activity with phenylacetaldehyde (a) and binds the screening substrate through hydrogen bonds and π-π stacking interactions. It is impossible for the substrate to bind in this conformation to the wild type protein due to Asn151 (b). The most active mutant with *S*-3-phenyl-1,2-propanediol is Val164Cys-Leu259Val-Cys362Gly (c), which binds the target substrate through hydrogen bonds to both hydroxyl groups and π-π stacking interactions. Wild type FucO cannot form as many hydrogen bonds and no π-π stacking interactions to the target substrate (d), which makes the enzyme inactive. Figure from publication I (Blikstad *et al.,* 2013).

5.1.4 Asn151 stabilizes enzyme complexes

In publication II, we continued the previous study and complemented it with a more in-depth structure-function analysis of active FucO mutants. The mutants of special interest, which were modeled and used for docking studies were: Asn151Gly, Phe254Ile, Leu259Val, Thr149Ser-Leu259Val, Val164Ile-Leu259Val, Thr149Ser-Asn151Gly-Leu259Val and Val164Cys-Leu259Val-

Cys362Gly, since these showed altered activity with either phenylacetaldehyde or *S*-3-phenyl-1,2-propanediol. This further highlighted the fact that mutation of Asn151 leads to activity with phenylacetaldehyde, because the experimental studies showed that also the other Asn151 mutants (Asn151Gly-Leu259Val and Thr149Ser-Asn151Gly-Leu259Val) had reasonable or high activity with the screening substrate. Furthermore, in accordance with the conclusions drawn in publication I, all Asn151 mutants showed no activity with the target substrate *S*-3-phenyl-1,2-propanediol. Moreover, the experimental results pinpointed a loss of activity with the natural substrate *S*-1,2-propanediol for the same mutants. Docking of the natural substrate to wild type FucO indicates a similar hydrogen-bonding network as for the target substrate *S*-3-phenyl-1,2-propanediol, with one hydroxyl group involved in formation of a hydrogen bond to the NAD^+ cofactor, while both hydroxyl groups interact with Asn151 (Figure 6a). Consequently, deletion of this residue will destabilize the *S*-1,2-propanediol – enzyme complex, which leads to the inability of the enzyme to perform catalysis. However, since Asn151 limits the size of the active site, mutagenesis of this residue was expected to install activity with the bulkier phenyl-substituted substrates, but as can be seen from the results with the natural substrate and the target substrate *S*-3-phenyl-1,2-propanediol, a more refined fine-tuning of the active site is needed for FucO to accept the bulkier substrates and perform proper catalysis.

5.1.5 Subtle changes install activity with the target substrate

In accordance with the conclusion that fine-tuning is needed, the studies show that already subtle changes are enough to install the wanted activity: the Leu259Val mutant is active with the target substrate *S*-3-phenyl-1,2-propanediol although the volume addition is only 25 $Å^3$ (Counterman & Clemmer, 1999), while the added phenyl ring on the target substrate would require ~100 $Å^3$ more compared to the natural substrate. The docked pose of *S*-3-phenyl-1,2-propanediol in the active site of the Leu259Val mutant shows hydrogen bonds between both hydroxyl groups on the target substrate and the cofactor (Figure 6b). Additionally, one hydroxyl group forms a hydrogen bond to Asn151. In wild type FucO, Leu259 creates a hydrophobic environment in the active site pocket and this, together with the sterical hindrance, repels one of the hydrophilic hydroxyl groups on the target substrate. The Leu259Val mutation does not reduce the hydrophobicity, but it creates enough space for the hydroxyl group to form favorable interactions with the cofactor instead. Hence, the size of the active site is an important factor, but the key to a successful biocatalyst is the ability to form a proper and stabilizing interaction network and, thereby, create a successful complex for catalysis. Consequently, the mutations have to be carefully selected not to interfere with the crucial hydrogen-bonding network taking place between *S*-3-phenyl-1,2-propanediol, Asn151 and the NAD^+ cofactor.

5.1.6 Val164 and Phe254 are involved in cofactor binding

The Val164Ile and Phe254Ile mutants caught special interest when tested with the natural substrate due to a two- and four-fold increase in k_{cat}, respectively. It is previously known that the release of NADH is the rate-limiting step for the FucO reaction (Blikstad & Widersten, 2010), which suggests that the cofactor dissociation rate is elevated in these mutants. Docking of the natural substrate to the Phe254Ile mutant shows no difference in binding mode compared to wild type FucO. However, both Val164 and Phe254 are interacting with the cofactor within a 4 Å distance, but upon the introduction of Phe254Ile, the π-π stacking interaction between the two ring systems is lost, which makes the cofactor less tightly bound and, therefore, the rate limiting dissociation reaction becomes faster (Figure 6c). Furthermore, the Phe254I mutant is not able to use *S*-3-phenyl-1,2-propanediol for catalysis, which can also be explained by the loss of the stabilizing π-π stacking ability. On the other hand, the effect of the Val164Ile mutation can possibly be accounted for by the introduction of a bigger and longer amino acid, which causes steric constraints for the cofactor. However, the increase in the reaction rate seen in both mutants can possibly be utilized to enhance the activity of other weakly performing mutants by engineering the binding site for the nucleotide.

5.1.7 Thr149 is important for side chain packing

Although Thr149 is far away from the cofactor and the active site, our results indicated that it is important for the catalytic function. In related proteins, this residue is either conserved or conservatively substituted so that the hydrogen-bonding capacity is retained and, in accordance, only the enzyme variants with a retained Thr149 or a conservative Thr149Ser mutation were able to have an appreciable activity with *S*-1,2-propanediol. Furthermore, Thr149 also affects the activity with *S*-3-phenyl-1,2-propanediol, since the Thr149Ser-Leu259Val mutant is 3.5-fold less active with the target substrate compared to the Leu259Val mutant (Figure 6d). Upon structural inspection, wild type FucO shows a 3.2 Å hydrogen bond between Asn151 and the hydroxyl group of Thr149, while it ranges between 2.7 Å and 5.0 Å in the set of ten Thr149Ser models. In wild type FucO, the β-methyl group of Thr149 restricts the rotational freedom of the side chain by interacting with Phe254 but since Ser lacks this methyl group, the side chain packing is affected and renders the Ser residue more flexible. In turn, this flexibility destabilizes the formation of a hydrogen bond between Ser149 and Asn151, which also increases the flexibility of the latter. Consequently, the hydrogen bonds between the important Asn151 and the substrate become labile, which negatively affects the catalysis. In short, effective catalysis requires both the hydrogen-bonding capacity of a hydroxyl group at position 149 and the right spatial position of this residue through proper side chain packing.

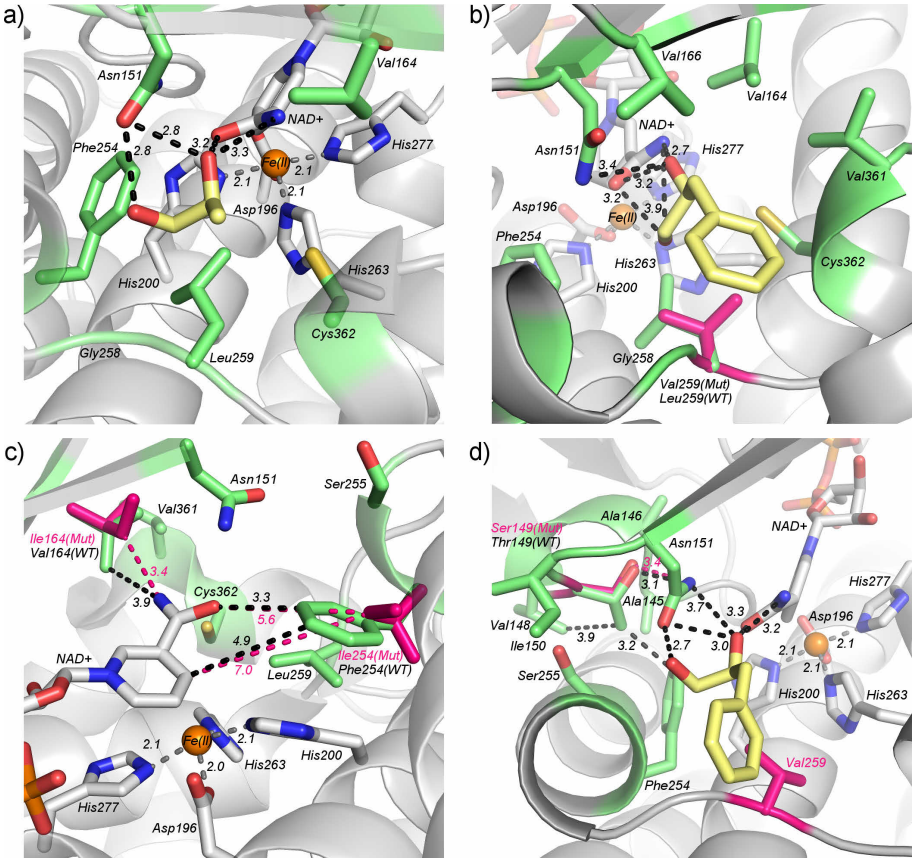


Figure 6. Amino acids important for FucO activity. All Asn151Gly mutants show a loss of activity with the target substrate *S*-3-phenyl-1,2-propanediol and the natural substrate *S*-1,2-propanediol. The docked pose of *S*-1,2-propanediol forms a hydrogen bond to the cofactor through one hydroxyl group, while both hydroxyl groups hydrogen bond to Asn151 in wild type FucO (a). This makes Asn151 important for the formation of a stable complex for catalysis. Wild type FucO is not active with *S*-3-phenyl-1,2-propanediol, but already a Leu259Val mutation installs activity with the target substrate. The docked pose of *S*-3-phenyl-1,2-propanediol in the active site of Leu259Val FucO, shows that the target substrate forms hydrogen bonds to the cofactor through both hydroxyl groups and one of them also interacts with Asn151 (b). The Leu259Val mutation is believed to create enough space to allow the second hydroxyl group to bind into the active site and form the hydrogen bonds. The Val164Ile and Phe254Ile mutants have higher turnover numbers, although they do not affect the binding of the natural substrate. However, the Val164Ile mutation reduces the space available for the cofactor, while the Phe254Ile mutation causes a loss of π-π stacking capacity (c). Both of these make the cofactor less tightly bound and consequently the dissociation rate is elevated. The Thr149Ser-Leu259Val mutant is less active compared to the Leu259Val mutant due to a more flexible Ser residue. This affects the side chain packing and renders Asn151 more flexible, which destabilizes the complexes (d). Figure adapted from publication II (Blikstad *et al.,* 2014).

5.1.8 FucO evolves through a generalist to a new specialist

Overall, the modeling and docking studies show that the use of FucO as a biocatalyst for substrates with a phenyl ring is limited by the active site cavity size, while hydrogen-bonding and π-π stacking capacity of several amino acids in the active site are important for stabilization of an enzyme-substrate complex. Usually, generalist enzymes with activity towards multiple substrates are considered to be more suited for biocatalyst production (Tracewell & Arnold, 2009). FucO is not a generalist enzyme but, as can be seen for other specialists (Matsumura & Ellington, 2001; Rockah-Shmuel & Tawfik, 2012), it gains specificity towards new substrates by first becoming a more promiscuous generalist enzyme (Leu259Val) and then adopting new specificity when the generalist version is mutated. The Leu259Val mutant is not a good catalyst but it shows activity with most of the substrates. However, the Asn151Gly-Leu259Val mutant is a new specialist enzyme with substrate specificity for phenylacetaldehyde. Also the mutants active with the target substrate *S*-3-phenyl-1,2-propanediol have the Leu259Val mutation coupled to other mutations, but their features resemble more the generalist enzyme. They are relatively inefficient; however, they do display activity with the target substrate and, hence, represent important intermediates, which can be further mutated into new specialist enzymes with high efficiency catalysis of *S*-3-phenyl-1,2-propanediol.

5.2 LpxR (*Y. enterocolitica*)

5.2.1 Introduction

Gram-negative bacteria have a protective outer membrane, which consists of LPS built from an O-antigen, a negatively charged core oligosaccharide and a hydrophobic membrane anchor called lipid A (Raetz, 1996; Raetz, 1990; Rietschel *et al.*, 1994). The lipid A component also serves as a key player in host-microbe interactions (Raetz *et al.*, 2007; Raetz *et al.*, 2009) and triggers innate immune system responses in mammalian cells upon encounter with LPS. The lipid A structure was earlier thought to be static, but it is now known to be modified by addition or deletion of fatty acids, phosphates, aminoarabinose, phosphoethanolamine and other decorations (Raetz *et al.*, 2007). The modifications may cause changes in the bacterial outer membrane physiology, as well as affect the biological activity of lipid A so that pathogens can avoid detection by the immune system of the host organism (Murata *et al.*, 2007; Trent *et al.*, 2006). For example, *Helicobacter pylori* and *Salmonella enterica* serovar *typhimurium* use this mechanism at the infection stage. These pathogenic bacteria employ the protein LpxR to deacetylate the 3' position of lipid A (catalysis presented in Figure 1a in Rutten *et al.,* 2009), which alters the recognition by the host cell immune system and, therefore, causes less inflammatory response (Kawasaki *et al.*,

2012; Reynolds *et al.*, 2006; Stead *et al.*, 2008). A corresponding lipid A species exists in *Y. enterocolitica* grown at 37 °C (Aussel *et al.*, 2000; Bengoechea *et al.*, 2003; Kawahara *et al.*, 2002; Oertelt *et al.*, 2001; Perez-Gutierrez *et al.*, 2010; Rebeil *et al.*, 2004; Rebeil *et al.*, 2006) and, furthermore, it has been shown that *Y. enterocolitica* virulence factors and lipid A structure are temperature dependent, with specific levels of acetylation and decorations at different temperatures (Marceau, 2005; Rebeil *et al.*, 2004; Reines *et al.*, 2012; Straley & Perry, 1995). The virulence factors aid in food borne infections of *Y. enterocolitica* in humans and animals, so that the bacteria can resist host cell defense mechanisms and colonize the intestinal tract (Bottone, 1997; Marceau, 2005; Straley & Perry, 1995). In publication III, we set out to determine whether *Y. enterocolitica* encodes for an LpxR ortholog, which would be responsible for the *Y. enterocolitica* lipid A species and, if so, to characterize this protein.

5.2.2 YeLpxR removes the 3'-acyloxyacyl residue from lipid A

The experimental results obtained by our collaborators showed the presence of a lipid A species lacking the 3'-acyloxyacyl residue in *Y. enterocolitica* and genome analysis confirmed the possibility of an LpxR ortholog. Deletion of the *lpxR* gene verified that the LpxR protein indeed is the lipid A 3'-*O*-deacylase in *Y. enterocolitica* (YeLpxR). However, this deacylation is more evident at 37 °C than at 21 °C, although the enzyme is expressed in higher levels in bacteria grown at the lower temperature. A general lack of function was ruled out as an explanation but, instead, the experimental results verified that the lipid A species found at 21 °C is decorated with aminoarabinose, which might inactivate YeLpxR or inhibit the physical interaction between the aminoarabinose-decorated lipid A and the enzyme. To explore these possibilities, a homology model of YeLpxR was constructed.

5.2.3 Asp31 is a key residue for YeLpxR substrate specificity

Amino acids 1-296 (signal sequence excluded) of YeLpxR (UniProtKB code A1JP43) were modeled based on the crystal structure of *Salmonella typhimurium* LpxR (StLpxR) (PDB code 3FID) (Rutten *et al.*, 2009), which shares 75 % sequence identity to YeLpxR. All amino acids determined to be important for StLpxR activity (Rutten *et al.*, 2009) are conserved in YeLpxR, but StLpxR is active also at 21 °C and, therefore, catalyzes lipid A decorated with aminoarabinose, which YeLpxR is unable to do. This indicates differences in substrate specificity despite the high sequence identity and the conserved functional amino acids. The homology model shows a reliable β-barrel fold, which is further supported by the high sequence identity to the crystallized StLpxR (Figure 7). Six amino acids differ in the active site of YeLpxR compared to StLpxR and two of these are major differences: a negatively charged Asp31 and a polar Gln35 in YeLpxR replace a small,

flexible Gly31 and a small, hydrophobic Ala35, respectively, in StLpxR (Figure 7). Of these, Gln35 is in the periphery of the YeLpxR active site, while Asp31 is in the center.

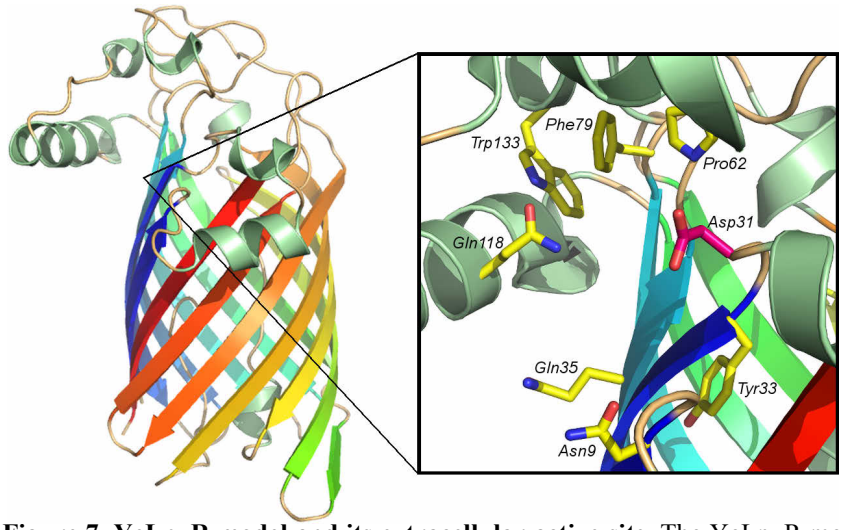


Figure 7. YeLpxR model and its extracellular active site. The YeLpxR model (left) shows a reliable β-barrel fold with intracellular turns and extracellular helices and loops. The active site residues (right, yellow sticks) show two major differences compared to the template StLpxR (PDB code 3FID; Rutten et al., 2009): Gln35 and Asp31, of which Asp31 (pink sticks) is located in the middle of the active site. Figure adapted from publication III (Reinés *et al.,* 2012).

Due to the big side chain, Asp31 forces a conserved Lys67 to adopt a slightly different conformation in YeLpxR compared to StLpxR and cavity calculations confirm that the volume of the active site cavity in StLpxR is bigger than in YeLpxR (Figure 8a). The changed Lys67 conformation in YeLpxR causes a loss of an inward protruding cavity next to this residue, while Asp31 also cuts the active site cavity in two parts with only a narrow connection. Hence, Asp31 causes major spatial limitations in the YeLpxR active site cavity. In order to explore the effect of an aminoarabinose decoration on the lipid A – YeLpxR interaction, modified Kdo_2-lipid A with and without aminoarabinose was docked to the homology model of YeLpxR and, for comparison, also to the crystal structure of StLpxR (PDB code 3FID) (Rutten *et al.*, 2009). When the lipid A molecule without aminoarabinose was docked to YeLpxR, the 4' phosphate group, which attaches aminoarabinose to lipid A, binds in the near vicinity of Lys67 and Asp31 (Figure 8b). On the other hand, when this molecule is docked to StLpxR, the same phosphate

group binds in the inward protruding cavity next to Lys67, which is lost in YeLpxR due to Asp31 (Figure 8c). As a result, the 4' phosphate is able to form more favorable electrostatic interactions with Lys67 in StLpxR compared to YeLpxR. As was expected based on the experimental results, the docking of the aminoarabinose-containing lipid A molecule to YeLpxR did not show any good and reliable results, but when docked to StLpxR, the aminoarabinose binds close to Gly31 and occupies the space corresponding to the narrow connection between the two cavities separated by Asp31 in YeLpxR (Figure 8d). This narrow connection is too small to accommodate the aminoarabinose decoration in the YeLpxR active site. Conclusively, the modeling and docking results indicate that lipid A modified with aminoarabinose at 21 °C simply cannot fit into the active site of YeLpxR due to the spatial restraints caused by Asp31 and, therefore, YeLpxR is unable to perform its function at this temperature and becomes latent. To verify this, YeLpxR mutants were constructed by site-directed mutagenesis and, indeed, the Asp31Gly mutant, which resembles StLpxR, deacylated aminoarabinose-containing lipid A in bacteria grown at 21 °C. Nevertheless, it is still possible that also other residues in YeLpxR have an effect on the enzyme latency. For example, Salmonella PagL was shown to become active by mutation of amino acids in extracellular loops, which suggests that they might be involved in recognition of lipid A decorated with aminoarabinose (Kawasaki *et al.*, 2005; Manabe & Kawasaki, 2008; Manabe *et al.*, 2010). Hence, similar regulation mechanisms might also play a role in YeLpxR latency.

5.2.4 YeLpxR helps the low inflammatory response upon infection

Additional experimental results showed that deletion of the YeLpxR enzyme makes the bacteria less motile and invasive, indicating that YeLpxR has an important role at the host colonization stage. Furthermore, the deletion of LpxR did not affect the production of the virulence factors with an anti-inflammatory effect. This means that *Y. enterocolitica* can employ both the anti-inflammatory virulence factors and lipid A 3'-*O*-deacylation to avoid responses from the host immune system and generate the characteristic low inflammatory response seen in *Y. enterocolitica* infections. Furthermore, by keeping the YeLpxR enzyme in a latent state at 21 °C, the bacteria can quickly respond and start the lipid A 3'-*O*-deacylation when entering the 37°C host cells. The importance of LpxR at the infection stage makes it an interesting drug target and the obtained results can be valuable for the design of new therapeutics against *Y. enterocolitica* infections.

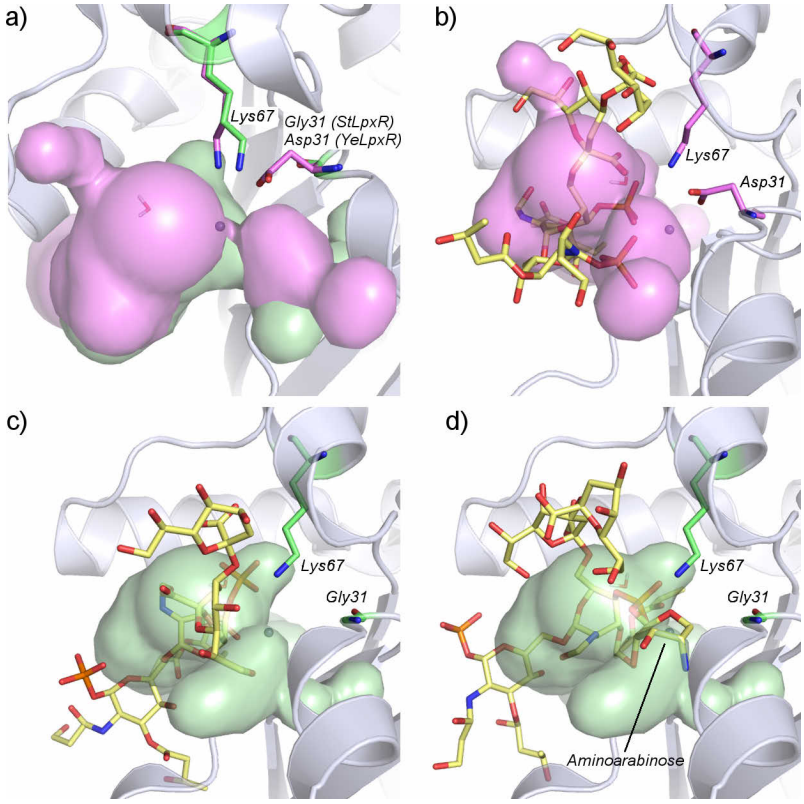


Figure 8. Docking of lipid A with and without aminoarabinose to YeLpxR (pink) and StLpxR (green). The StLpxR active site (green) is bigger than the YeLpxR active site (pink), which is essentially cut in two parts due to Asp31 (a). Both YeLpxR (b) and StLpxR (c) catalyze lipid A (yellow sticks) without aminoarabinose, but only StLpxR is able to use aminoarabinose-containing lipid A (d). The aminoarabinose binds in between Lys67 and Gly31 at the same place where Asp31 in YeLpxR cuts the active site in two parts, thereby making YeLpxR unable to accommodate and catalyze aminoarabinose-containing lipid A. Figure from publication III (Reinés *et al.,* 2012).

5.3 LpxO (*K. pneumoniae*)

5.3.1 Introduction

K. pneumoniae is another bacterium employing lipid A modifications to avoid the host immune system and uses the LpxO protein (KpLpxO) for it. *K. pneumoniae* is a pathogenic bacterium, which upon infection significantly affects the blood stream and the respiratory functions in humans and

increasing rates of multidrug resistance has become a problem (De Majumdar *et al.*, 2015). LpxO is a relatively unknown protein but shows homology to the catalytic domain of bovine Asp/Asn β-hydroxylase, which is a Fe^{2+}/α-ketoglutarate-dependent dioxygenase (Gibbons *et al.*, 2000). KpLpxO has 302 amino acids with a hydrophobic N- and C-terminus, it displays conserved His, Asp and Glu residues, which have been shown to be important in bovine Asp/Asn β-hydroxylase (McGinnis *et al.,* 1996) and, in particular, KpLpxO has a conserved Fe^{2+} binding motif (His-X-Asp-$(X)_{\sim 50}$-His). Bacterial LpxO homologs are all of similar length and closely related to each other (Raetz, 2001), which indicates a similar function. It has been shown *in vitro* that KpLpxO modify the lipid A at position 2' by adding a hydroxy-myristate group (Clements *et al.*, 2007; Llobet *et al.*, 2011) (catalysis presented in Figure 2a in Gibbons *et al.*, 2008). In publication IV, we wanted to analyze the lipid A species *in vivo* and structurally characterize KpLpxO to find important amino acids for its function.

5.3.2 KpLpxO is involved in 2-hydroxylation of lipid A

The experimental mutation of the *lpxO* gene showed that *K. pneumoniae* encodes for the enzyme LpxO (UniProtKB code W9B4N2), which has been implicated in 2-hydroxylation of lipid A. This mutant is unable to hydroxylate C_{14} on the primary 2'-linked R-3-hydroxymyristoyl group, while a plasmid containing the *lpxO* gene restores this hydroxylation when cloned into *K. pneumoniae*. Furthermore, the mutant strain lacking LpxO was found in lower amounts in trachea and lung compared to wild type *K. pneumoniae* and, thereby, demonstrates that the 2-hydroxylated lipid A modification helps *K. pneumoniae* avoid the innate immune system and attenuate the inflammatory responses of *K. pneumoniae*. Also, the effect of antimicrobial peptides, especially colistin, which is one of the few options left to treat multiresistant *K. pneumoniae* infections, is counteracted by this lipid A modification.

5.3.3 KpLpxO adopts the Asp/Asn β-hydroxylase fold

3D structural modeling of KpLpxO showed that it contains transmembrane helices in both the N- and C-terminus, with a central, cytoplasmic domain from amino acid 19 to 279. This central domain contains the active site and a model could be based on the crystal structure of human Asp/Asn β-hydroxylase (PDB code 3RCQ) (Krojer *et al.,* to be published). Due to low sequence identity, KpLpxO was aligned with homologs in a multiple sequence alignment, after which the sequence for human Asp/Asn β-hydroxylase was aligned. Comparison of KpLpxO to the well characterized Asp/Asn β-hydroxylase from bovine (75 % sequence identity to human Asp/Asn β-hydroxylase) (McGinnis *et al.*, 1996) shows that several catalytically important amino acids are conserved in KpLpxO and, together

with the resulting model, it indicates that KpLpxO indeed adopts the Asp/Asn β-hydroxylase fold and has the typical iron-binding motif His-X-Asp-$X_{\sim 50}$-His in the active site (His_{155}-Arg-Asp-X_{44}-His_{202} in KpLpxO) (Figure 9). From the model, we could deduce the amino acids within 4 Å from the predicted active site and then experimentally mutate these to alanine to verify their importance. The mutated residues were His155, Asp157, Arg164, His166, Trp188, Glu198, His202, Arg212 and Asp218, of which all, except His166, are strictly conserved. The mutants were not able to produce 2-hydroxylated lipid A although the mutants of KpLpxO were produced. Hence, our results show that KpLpxO is likely to adopt the Asp/Asn β-hydroxylase fold and that the active site and catalytic residues are similar to bovine Asp/Asn β-hydroxylase. Mutation of His155, Asp157 and His202 were expected to have an effect since these residues are involved in the binding of iron. Furthermore, mutation of Arg163, His166 and Arg212 was also highly likely to affect the catalysis since these residues are pointing straight into the active site and probably are involved in interactions with the substrate. On the other hand, Trp188, Glu198 and Asp218 can have a structurally important role rather than being directly involved in catalysis. Trp188 could be keeping Arg212 in the right conformation for catalysis, while Glu198 and Asp218 could have a similar role for the correct positioning of Arg164. However, these residues were all confirmed to be important for the proper function of KpLpxO and the results lead the way for more detailed structural characterization of this protein and the design of new drugs against *K. pneumoniae* infections.

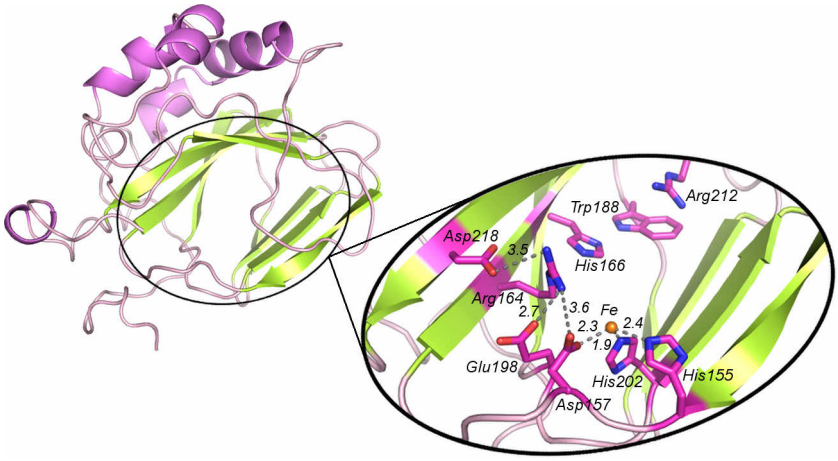


Figure 9. Structural model of KpLpxO and active site. KpLpxO contains a central soluble domain, which adopts the Asp/Asn β-hydroxylase fold with an iron-binding motif (His_{155}-Arg-Asp-X_{44}-His_{202}) and several catalytically important amino acids (pink sticks). Figure from publication IV (Llobet *et al.,* accepted manuscript).

5.4 Slr0006 (*Synechocystis*)

5.4.1 Introduction

Synechocystis is a carbon concentrating cyanobacterium, which up-regulates multiple genes coding for proteins with unknown function when subjected to low CO_2 and, hence, a limited access to inorganic carbon (Battchikova *et al.*, 2010). One of these genes is *slr0006* (Carmel *et al.*, 2011), which encodes for the 23 kDa Slr0006 protein (UniProtKB code Q55667). So far, this cytosolic protein does not seem to be important for cell survival (Carmel *et al.*, 2012). However, the cellular localization shifts upon addition of divalent cations, which link the Slr0006 protein to the membrane (Carmel *et al.*, 2011). Our aim in publication V was to functionally characterize the Slr0006 protein and couple the results to structural analyses.

5.4.2 Slr0006 belongs to the Sua5/YciO/YrdC protein family

Sequence analysis of Slr0006 indicates that it belongs to the Sua5/YciO/YrdC family of proteins, which is one of the top 10 universal protein families with an unknown function (Galperin & Koonin, 2004). However, several members, such as *E. coli* YrdC (PDB code 1HRU) (Teplova *et al.*, 2000), *E. coli* YciO (PDB code 1KK9) (Jia *et al.*, 2002), *Sulfolobus tokodaii* Sua5 (C-terminal domain) (PDB code 2EQA) (Agari *et al.*, 2008), *Methanothermobacter thermoautotrophicum* Mth1692 (PDB code 1JCU) (Yee *et al.*, 2002), *E. coli* HypF (middle domain) (PDB code 3TSQ) (Petkun *et al.*, 2011) and *Streptococcus mutans* smu. 1377c (PDB code 3L7V) (Fu *et al.*, 2010) have a known crystal structure. Slr0006 shares a low sequence identity to these proteins: 19.0 % to the C-terminal YrdC-domain of Sua5, 18,3 % to YrdC and 17 % to YciO. However, the low sequence identity is a common feature for the Sua5/YciO/YrdC protein family, although the proteins share a highly similar structure. For example, despite only 27 % identity, YciO and YrdC share a similar structure and also the smu. 1377c protein has a similar fold as YrdC, YciO and the C-terminal domain of Sua5, although the sequence identity is 15 %, 25 % and 16 %, respectively. Furthermore, structures are three to ten times more conserved than sequences (Illergard *et al.*, 2009), which is well represented by the other members of the Sua5/YciO/YrdC protein family and indicates a similar behavior for Slr0006. Due to the low sequence identity, a multiple structure-based alignment was calculated by superimposing YciO, the YrdC domain of Sua5, YrdC and Mth1692, and the Slr0006 sequence was then aligned to the pre-aligned multiple structure-based sequence alignment. Three models were created: amino acids 1-211, 10-193 and 1-206 of Slr0006 were modeled based on the C-terminal domain of Sua5, YrdC and YciO, respectively. The 3D models of Slr0006 show that it is possible for this protein to adopt the Sua5/YciO/YrdC fold with an α/β twisted open-sheet structure, containing both parallel and

anti-parallel β-strands (Figure 10a). The model evaluation also shows that the models are of good quality and reliable.

5.4.3 Slr0006 could bind RNA or nucleotides

Another feature that is shared among all crystal structures of the Sua5/YciO/YrdC family and the Slr0006 models is a central cavity with a strong positive charge (Figure 10b). This cavity binds RNA in YrdC (Teplova *et al.*, 2000) and, therefore, Slr0006 was experimentally tested for association with the protein synthesis machinery, *i.e.* ribosomes. The results show that Slr0006 always co-localizes with the S1 protein of the 30S ribosomal subunit, which implies a possibility for RNA-binding. Moreover, both YrdC and Sua5 are highlighted as essential for N^6-threonylcarbamoyl adenosine (t^6A) biosynthesis (El Yacoubi *et al.*, 2009) and Sua5 has been crystallized in complex with the essential biosynthesis components L-threonine and an ATP analogue (ANP) (Kuratani *et al.*, 2011). The t^6A biosynthesis pathway exists in all organisms with a sequenced genome and they also encode one or more members of the Sua5/YciO/YrdC family of proteins (El Yacoubi *et al.*, 2009). However, the *Synechocystis* genome encodes the Sll1866 protein, which corresponds better to both the C-terminal YrdC domain of Sua5 and YrdC itself: Sll1866 has 27 % identity to the C-terminal domain of Sua5 and 29 % identity to YrdC, compared to 20 % and 19 %, respectively, for Slr0006. This indicates that Sll1866, rather than Slr0006, would have a similar function to *E. coli* YrdC.

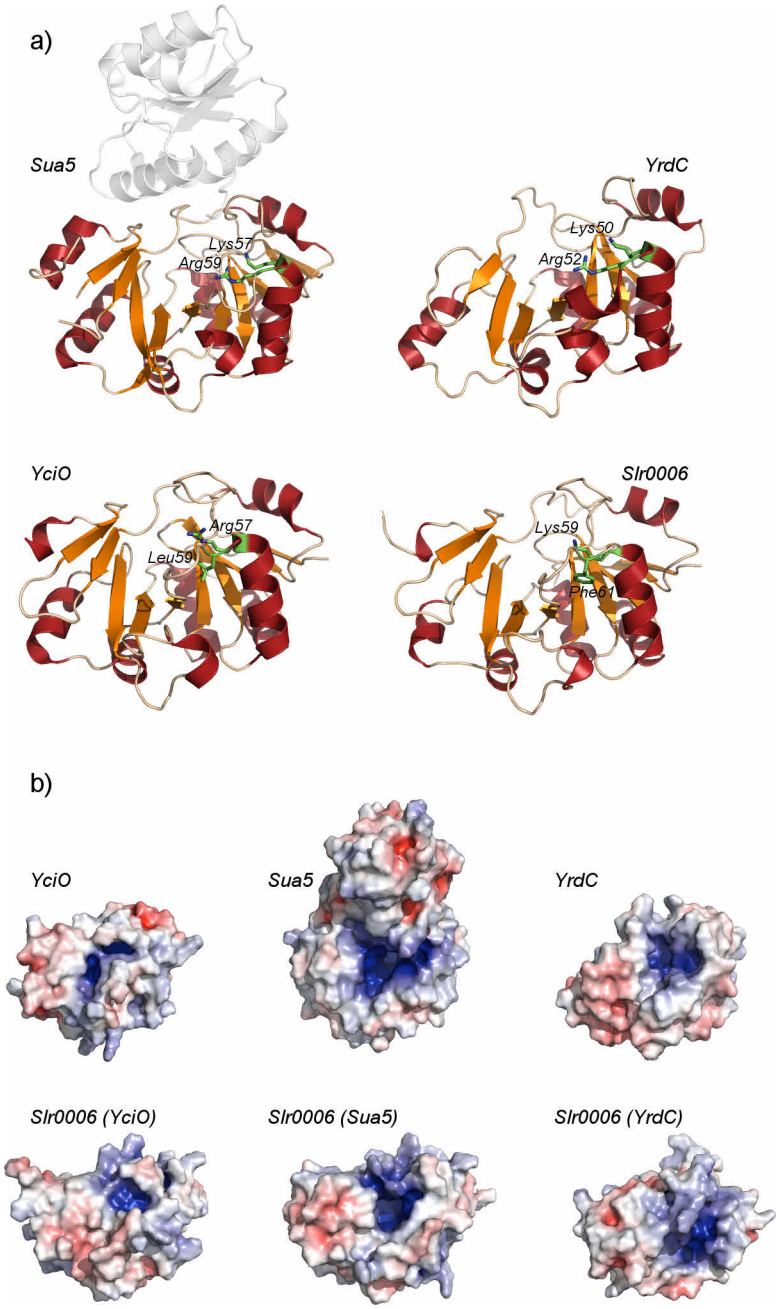


Figure 10. The Sua5/YciO/YrdC fold and electrostatic surface. a) shows the overall fold of the Sua5/YciO/YrdC family in comparison to the Slr0006 model based on YciO. b) shows the electrostatic surface of the Sua5/YciO/YrdC proteins in comparison to the models of Slr0006 (red are negatively charged areas, blue are positively charged, grey are neutral and color ranges from -7 to 7). Figure from publication V (Carmel *et al.,* 2013).

5.4.4 Slr0006 belongs to the YciO family

The YciO protein has been shown not to be a functional ortholog to YrdC, although it adopts the YrdC architecture. Hence, El Yacoubi *et al.* (2009) suggested that YrdC and YciO should be split in two families. The dividing factor would be the two positively charged residues Lys50 and Arg52 according to YrdC numbering. These are conserved in most of the Sua5/YciO/YrdC family members, but YciO has the positively charged Arg57 corresponding to Lys50 (YrdC), while a hydrophobic Leu adopts the position corresponding to Arg52 (YrdC) (Figure 11). Slr0006 follows the YciO pattern rather than the one for YrdC: Lys59 (Slr0006) corresponds to Lys50 (YrdC), while the hydrophobic Phe61 (Slr0006) replaces Arg52 (YrdC). Furthermore, based on the analyses by Petkun *et al.,* (2011), several amino acids are conserved in YrdC-like proteins (Arg245, Ala/Ile251, Lys/Phe294 and Asn324 [HypF numbering]) but different in YciO (Leu, Phe, Lys, Met, respectively), which indicates a different substrate for YciO than for the YrdC-like proteins. Similarly, five of the HypF nucleotide-binding residues (Arg245, Ala/Ile251, Thr321, Asn324 and Val/Ile363) differ in Slr0006 (Phe, Leu, Ala, Lys and Leu, respectively), which suggests that Slr0006 resembles YciO more than YrdC and, therefore, probably binds similar ligands to YciO.

Although the function of YciO is still unknown, the protein has been implicated in glycogen metabolism (Montero *et al.*, 2009), since *E. coli* lacking the *yciO* gene accumulated enormous amounts of glycogen. This led us to test the same for the mutant lacking the *slr0006* gene. In contrast, this mutant showed similar amounts of glycogen for both wild type and the mutant, implicating that the functions of Slr0006 and YciO are different, although the structure is similar. A BLAST search against the *Synechocystis* genome with the sequence for YciO as query then showed that the Sll0216 protein has a 40 % sequence identity to *E. coli* YciO, which indicates that Sll0216 performs the same function as YciO rather than Slr0006. Despite this, the co-localization of Slr0006 and the S1 protein of the 30S ribosomal subunit indicates a role in processes related to ribosomes, but this function is probably unique compared to the characterized Sua5/YciO/YrdC family members, since the important amino acids are not conserved. However, the conserved positively charged cleft on the surface of Slr0006 indicates that the function can be related to nucleotide- or RNA-binding.

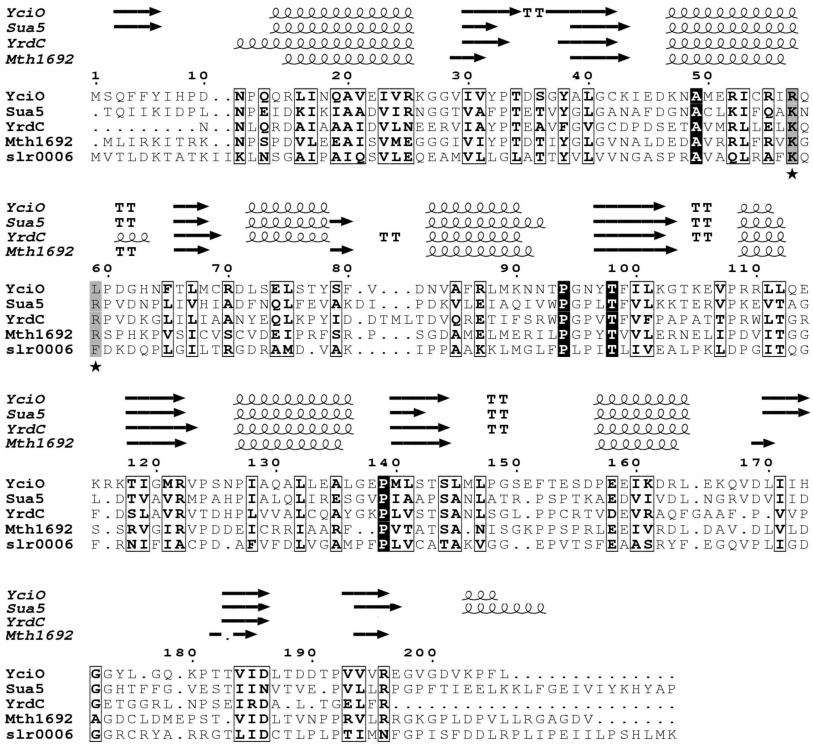


Figure 11. Structure-based alignment of Slr0006 and crystallized members of the Sua5/YciO/YrdC family. The residues used to divide YciO from YrdC are marked with grey boxes and black stars. Black boxes show conserved residues, bold letters similar residues. The secondary structure of crystallized Sua5/YciO/YrdC members are shown on top of the alignment. Figure adapted from publication V (Carmel *et al.,* 2013).

5.4.5 Slr0006 could contribute to a bigger complex

The YrdC domain is also found fused to other domains in multidomain proteins, such as the nucleotide binding proteins HypF (Petkun *et al.*, 2011; Teplova *et al.*, 2000) and TobZ (Parthier *et al.*, 2012), which have the YrdC domain coupled to Kae1-like domains. In turn, Kae1 proteins are part of the Kae1/Qri7/YgjD family, which is also universally conserved and participates in the t^6A biosynthesis (El Yacoubi *et al.*, 2011; Galperin & Koonin, 2004). However, the function of the proteins in the Kae1/Qri7/YgjD family has been suggested to differ from the proteins in the Sua5/YciO/YrdC family so that they together represent the whole synthesis pathway (El Yacoubi *et al.*, 2011). If this is the case, there is still one unidentified and uncharacterized protein in the pathway and, therefore, it can be speculated that Slr0006 would be similar to this protein. Furthermore, it can be hypothesized that Slr0006

could be part of a bigger protein complex, since members of the Kae1/Qri7/YgjD family have been indicated to contribute to a bigger hetero-oligomeric enzyme in bacteria. In these types of enzymes, the subunits can be homologs, which in many cases have lost their catalytic function and serve as interaction surfaces. Hence, this could be the role of Slr0006 and, therefore, it might not be essential for the survival of the bacteria. These results highlight the complex mechanisms behind the adaptation of plants and cyanobacteria to new environmental factors, which can be caused by for example global warming.

5.5 CIP2A (Human)

5.5.1 Introduction

CIP2A is a 905 amino acid long oncoprotein, with a molecular mass of ~102 kDa (UniProtKB Q8TCG1). It can be used as a clinically relevant prognostic marker in most human cancer types (Khanna *et al.*, 2011) and has been shown to interact with multiple proteins functioning as substrates for protein phosphatase 2A (PP2A) dephosphorylation, such as MYC, E2F1, AKT, death-associated protein kinase 1 (DAPK1), and the rapamycin complex 1 (mTORC1) (Puustinen & Jäättelä, 2014). Especially the cancer promoting effect of the CIP2A-MYC interaction has caught interest. PP2A normally dephosphorylates MYC, which results in the degradation of MYC, but CIP2A can disrupt this chain of events by preventing the dephoshporylation and, thereby, stabilizing MYC (Junttila *et al.*, 2007). The elevated MYC levels then help the cells to transform and induce malignant cell growth. CIP2A is highly interesting as an anticancer drug target (He *et al.*, 2012; Khanna *et al.*, 2013) due to its overexpression in many cancer types, such as ovarian cancer (Bockelman *et al.*, 2011), breast cancer (Come *et al.*, 2009), non-small cell lung cancer (Dong *et al.*, 2011), gastric cancer (Khanna *et al.*, 2009), bladder cancer (Xue *et al.*, 2013), head and neck squamous carcinoma, colon cancer (Junttila *et al.*, 2007) and liver cancer (Soo Hoo *et al.*, 2002). However, the lack of structural data on CIP2A significantly hampers the drug development efforts since it is not even known if CIP2A is a druggable protein or not (Khanna *et al.*, 2013). Therefore, all structural data on this protein is essential for the development of new therapeutics targeting the cancer-causing mechanism of CIP2A. Hence, in publication VI, we set out to model the structure of CIP2A and find important amino acids for protein-protein or protein-ligand interactions, which could ultimately aid in the development of new anticancer drugs exerting their effect through CIP2A.

5.5.2 CIP2A N-terminus adopts the armadillo fold

Analysis of the CIP2A amino acid sequence with SMART (Letunic *et al.*, 2015; Schultz *et al.*, 1998) indicated that CIP2A consists of an armadillo repeat fold (ArmRP) from amino acid 47 to amino acid 308 (CIP2A-ArmRP), followed by a coiled coil region between amino acids 636 and 884. The ArmRP fold is characterized by a right-handed superhelix formed by four to twelve motifs of three α-helices with ~42 amino acids each (Coates, 2003; Reichen *et al.*, 2014). The secondary structure predictions agree with this by proposing an all α-helical profile for CIP2A-ArmRP and also SCOP suggests that the CIP2A-ArmRP is similar to the ArmRP protein β-catenin (PDB code 1JDH) (Graham *et al.*, 2001). Furthermore, the highly conserved sequence logo Leu-Val-X-Leu-Leu, deduced from naturally occurring and designed ArmRP proteins (Parmeggiani *et al.*, 2014) is also conserved in CIP2A-ArmRP, which further improves the reliability of a CIP2A-ArmRP domain.

BLAST search against PDB at the NCBI server did not result in any good templates for modeling of either full length CIP2A or CIP2A-ArmRP. Hence, the model was produced with 3D structure prediction servers, which employ highly sensitive methods to detect remote homologs, align the query protein to the resulting templates and produce a hypothetical 3D model (Söding, 2005). The predictions for full length CIP2A were not reliable when compared to the results from sequence analysis, since the servers predicted an ArmRP fold for the whole sequenc, while the sequence analysis indicated a C-terminal coiled coil domain. However, the model of CIP2A-ArmRP showed reasonably good quality. HHpred (Remmert *et al.*, 2011; Söding *et al.*, 2005; Söding, 2005), Phyre (Kelley & Sternberg, 2009) and I-TASSER (Roy *et al.*, 2010; Roy *et al.*, 2012; Zhang, 2008b) predicted the CIP2A-ArmRP domain to fold into a similar structure as the synthetic OR329 arm8 protein (PDB code 4HXT) (Parmeggiani *et al.*, 2014) and the CIP2A-ArmRP model produced by I-TASSER was used for further analysis, since I-TASSER has been proven to generate the best 3D structure predictions among all automated servers in the CASP 7-10 experiments (Zhang, 2014). Also, the model itself shows a good quality score and a good topological similarity to the template. Furthermore, several evaluation programs and visual inspection together with comparison to the template indicated that the model is of good quality and reliable. However, manual evaluation showed that CIP2A-ArmRP Arg229 was pointing into a hydrophobic environment inside the protein, which is uncommon for a charged amino acid unless they form salt bridges. Hence, we searched for different loop conformations with the Loopy program in Jackal and chose a low energy conformation, where Arg229 interacts with the solvent instead. Parallel MD simulations were performed for both the model and the crystal structure to verify a stable fold with no changes or unfolding during the simulation and the results confirmed

that the fold is stable with rigid secondary structure elements and no unfolding taking place during the simulation.

5.5.3 The central groove in CIP2A-ArmRP binds peptides

The CIP2A-ArmRP model has 18 α-helices, thereby forming six ArmRP motifs, from which the last α-helices (H3) create a central groove when the motifs twist into the right-handed superhelix (Figure 12). The ArmRP fold is structurally highly conserved even though the sequence identity between ArmRP proteins might be low. Furthermore, ArmRP proteins are rigid structures with a typical central groove (Coates, 2003; Reichen *et al.*, 2014; Varadamsetty *et al.*, 2012), which is known to be a binding site for peptides from bigger proteins (Cutress *et al.*, 2008; Eklof Spink *et al.*, 2001; Reichen *et al.*, 2014). A MetaPocket (Huang, 2009; Zhang *et al.*, 2011) search indicated that the central groove is indeed a binding site for CIP2A-ArmRP as well. This central pocket is made up of residues Gln122, Gln125, Met160, Pro161, Gly164, Asn168, Arg171, Val206, Phe207, Ser210, Ser213, Ser214, Leu217, Leu249, Lys252, Tyr253, Asp256, Met259, Asp260, which are all well conserved in multiple sequence alignments of homologous proteins. Met160, Pro161, Phe207, Ser213, Leu249, Lys252, Tyr253 and Met259 are more variable than the rest of the residues, while Asn168, Ser214, Leu217 and Asp256 are strictly conserved and, hence, probably important for structure and/or function of CIP2A-ArmRP.

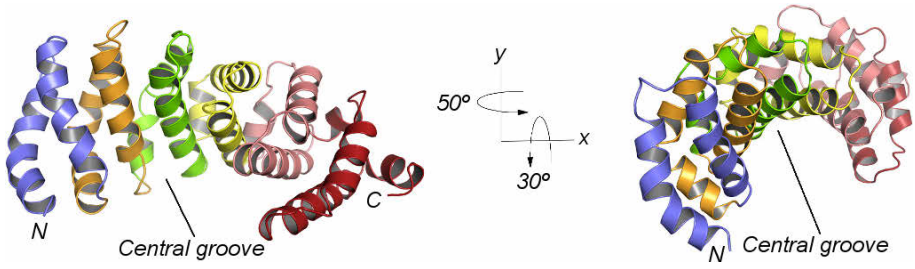


Figure 12. 3D structural model of CIP2A-ArmRP and its central groove. The model shows a reliable ArmRP fold, with six ArmRP repeats twisting around an axis to form a superhelix and a concave central groove. Figure from publication VI (Dahlström & Salminen, 2015).

I-TASSER also predicts a protein/peptide binding function for the central groove in CIP2A-ArmRP. All of the most similar proteins found in PDB by I-TASSER have a protein binding function, which indicates that also CIP2A-ArmRP would bind a peptide from a bigger protein in the central groove. Moreover, ConSurf analysis shows that the central groove is highly

conserved between CIP2A-ArmRP and homologs, further implicating this as an important binding site. The binding mode in naturally occurring ArmRP proteins is very conserved: the extended peptide is bound antiparallel to the ArmRP motifs into the central groove with hydrogen bonds between conserved Asn residues in the H3 α-helices and the peptide backbone to keep it in an extended conformation (Andrade *et al.*, 2001; Conti *et al.*, 1998; Graham *et al.*, 2001; Ishiyama *et al.*, 2010; Morishita *et al.*, 2011; Roman *et al.*, 2013; Tarendeau *et al.*, 2007). Specificity is conferred by other residues in the central groove, which interact with the side chains of the amino acids in the bound peptide (Reichen *et al.*, 2014). I-TASSER suggests, with high confidence, that CIP2A-ArmRP would bind a peptide in a similar manner as the human adenomatous polyposis coli (APC) protein fragment binds to mouse β-catenin (PDB code 1JPP) (Eklof Spink *et al.*, 2001). A polar ladder in β-catenin, which is conserved also in other ArmRP proteins (Andrade *et al.*, 2001), binds the APC protein fragment to the central groove in an extended conformation. In CIP2A-ArmRP, Gln82, Gln119, Gln122, Gln125, Gln311, Asn130, Asn168, Asn173, Asn218, Asn264 and His172 form a similar polar ladder and the high degree of conservation of these residues implicates that they are essential (Figure 13). Furthermore, we superimposed the CIP2A-ArmRP model on the APC-β-catenin complex and analyzed the residues within 4 Å of the bound peptide, which all coincide or are in the near vicinity of the polar ladder. Hence, CIP2A-ArmRP is highly likely to bind a peptide from an interaction partner in the same way as other ArmRP proteins.

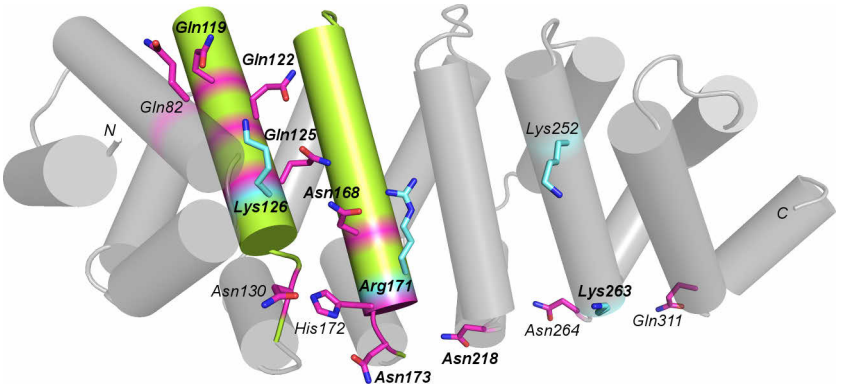


Figure 13. Polar ladder in CIP2A-ArmRP. The polar ladder in the CIP2A-ArmRP domain (pink sticks) is formed by highly conserved amino acids. These have an important function in peptide binding together with conserved positively charged residues (cyan sticks), which confer specificity to the bound peptide. Strictly and highly conserved residues are shown in bold. Figure from publication VI (Dahlström & Salminen, 2015).

5.5.4 CIP2A interaction partners have a conserved binding motif

It is previously known that specific motifs are common in the ArmRP interaction partners depending on the charge of the central groove in the ArmRP protein itself (Reichen *et al.*, 2014). The electrostatic surface calculations show a strong positive charge for the CIP2A-ArmRP central groove, while the opposite side of the protein is divided into one positively charged area and one negatively charged area. Also, the top and the bottom of CIP2A-ArmRP show clearly defined areas of positive and negative charge. Similarly to CIP2A-ArmRP, β-catenin has a positively charged groove and has been shown to interact with peptides containing a conserved Asp-X-Hp-Hp-X-Ar-X_{2-7}-Glu motif (X = any amino acid, Hp = a hydrophobic residue, Ar = aromatic residue), where the conserved Asp and Glu form salt bridges with two Lys residues in the central groove of β-catenin (Graham *et al.*, 2000; Sun & Weis, 2011; Xu & Kimelman, 2007). Similarly, CIP2A-ArmRP exhibits the highly conserved Arg171 and Lys263, along with the conserved Lys252 and Lys126, indicating that the interaction mode between CIP2A-ArmRP and a peptide would be highly similar to the interaction seen in peptide-β-catenin complexes. Furthermore, this means that it is likely that the peptide interacting with CIP2A-ArmRP would show the same conserved motif (Asp-X-Hp-Hp-X-Ar-X_{2-7}-Glu) as the peptides interacting with β-catenin and other ArmRP proteins with a positively charged central groove. PP2A (Junttila *et al.*, 2007), MYC, E2F1, the mTORC1 complex (De *et al.*, 2014), AKT, DAPK1 (Puustinen & Jäättelä, 2014) and H-Ras (Wu *et al.*, 2015) are all known to interact with CIP2A. However, analysis of the MYC, H-Ras and AKT sequences did not reveal the expected conserved motif, which indicates that these would interact with CIP2A outside of the CIP2A-ArmRP central groove. However, the 65 kDa scaffolding subunit (also called A or PR65 subunit) of PP2A, E2F1, DAPK1 and the DPTOR and RPTOR subunits of the mTORC1 complex all have the conserved motif in their sequences. Also the mTOR subunit of the mTORC1 complex has a similar motif, where an Asp replaces the conserved Glu but, despite this, the negative charge is conserved. Hence, all of these proteins can be expected to interact with the central groove in CIP2A-ArmRP. These results may be of considerable help when designing experiments to characterize the CIP2A function and also for future crystallization efforts. The results indicate a possibility to generate a construct for the CIP2A-ArmRP domain, which might be easier to crystallize than the whole protein due to the rigidity of the ArmRP fold. This could then give the much-needed information for development of anti-cancer drugs, which work through CIP2A.

5.6 Choosing between BLAST results

The resulting reliability of the structural model is already affected at the sequence/structure search step of the modeling procedure. There are several

sequence databases and, therefore, it is good to keep in mind to try to maximize the information obtained from them. When performing modeling studies, coupling the results to existing literature is of key importance to be able to draw biological conclusions based on the models. Therefore, when searching for homologs, start in smaller databases like UniProtKB, which is manually annotated and contains biological information about known signal sequences, structures, function data etc. If the results are not sufficient, then gradually increase the database size to obtain homologs, which may not be experimentally characterized or are predicted from sequencing projects. For example, the CIP2A protein (publication VI) required the large non-redundant sequence database to find homologs for multiple sequence alignment, while LpxO homologs were found in UniProtKB (publication IV). Hence, there was a possibility to check for experimental information on the LpxO homologs, while some of the CIP2A homologs were only predicted to exist. This type of knowledge is very valuable when interpreting the sequence alignment and determining amino acids important for function and/or structure. This further highlights the benefits of smaller, manually annotated databases, since these incorporate references and serve as a great starting point for gathering the necessary and valuable experimental information for the modeling procedure.

The search for similar sequences or structures is usually done through BLAST searches. The results are reported in table form with name of the protein, query coverage, statistical significance (E-value), sequence identity etc. At this point, it is crucial to evaluate especially query coverage, statistical significance (expected value or E-value) and sequence identity together and not base the choice of sequences or structures solely on sequence identity. The sequence identity might be sufficiently high for some of the results, which only represent a small part of the protein of interest and, therefore, are not valuable for the study. Moreover, it is important to pay attention to the E-value, which describes the number of hits that can be seen by chance. The lower this value is, the more relevant is the hit and a good rule of thumb says that the E-value should be below 0.001. Also other factors should be taken into account when choosing a good template for the modeling. In addition to the factors mentioned above, the resolution, the presence/absence of ligands or cofactors and mutations should be taken into account depending on the question of interest. Furhermore, conformational changes and missing parts in the structure can also affect the end result. It is also a good idea to compare a secondary structure prediction for the target protein to the secondary structure of the template protein to see if they correspond. Hence, it is not enough to consider a single value when choosing homologous sequences or structures to use as templates. Instead, all of these different values should be considered in a bigger picture and with the research question in mind.

5.7 Sequence alignment is the most important step

One major drawback of the automated servers is the inability to edit the sequence alignment. This is the most important step in the whole modeling process, directly affecting the quality and reliability of the resulting model and, therefore, it requires careful and thorough inspection. The ability to inspect and modify sequence alignments requires understanding of amino acids and their properties, but it is also essential to properly determine the type of alignment needed to increase the accuracy and reliability of the model. If the sequence identity is high, like for the LpxR proteins (publication III), global pairwise alignment can be used to produce a reliable model. However, low sequence identity between the target and the template protein requires a multiple sequence alignment to improve the reliability of the model, which is exemplified by the LpxO protein (publication IV). When performing a multiple sequence alignment, it is important to include a number of sequences with varying sequence identity to the target protein. If the sequence identity range is too narrow and only include sequences with high sequence identity, the high amount of conserved amino acids will pose problems when determining the amino acids that are truly important for the function and/or structure. If sequences with too low sequence identity are included in the alignment, they will introduce gap areas of different size and might differ too much from the other sequences for the important amino acids to be detectable. Hence, a proper range of sequence identity should be chosen for each specific protein but, usually, a range of 30 – 80 % sequence identity is appropriate.

Generally, the protein fold is more conserved than sequence (Illergard *et al.*, 2009), which is also true for Slr0006 (publication V), where a pairwise sequence alignment would not give a reliable result. Even multiple sequence alignment shows some unreliable gap areas, which is why the conserved fold was chosen as the basis for the alignment instead. By superimposing the known structures, a multiple structure-based sequence alignment could be generated and this produced a model with a more reliable fold than the other alignment methods would have. The aim of the study was to model the fold and find amino acids important for the function, which means that the obtained accuracy and reliability were sufficient for this project. On the other hand, the analyses performed on FucO (publication I and II) and LpxR (publication III) demanded more detailed information about active site amino acids, which required high accuracy models, especially for the active site, to give reliable results. The availability of a FucO crystal structure and a high sequence identity homolog with known structure for YeLpxR, ensured that the resulting models were highly reliable and accurate. However, if the overall sequence identity is low, it might be beneficial to tackle the alignment problem through local alignments with, for example, only active site residues. This enables high accuracy in the area of interest, while the less

important areas may remain unstudied. Hence, the purpose of the project and the aim of the analysis should always be kept in mind when aligning sequences for structural modeling and the way to produce an alignment and a model should always be malleable depending on the obtained results.

5.8 Considerations when creating the structural model

At the model creation step, several models should be created to get different conformations for the amino acid side chains and, hence, varying overall energies. Usually, a set of ten models is enough and often the model with the lowest energy is the best, but it is a good idea to check all the models visually to see the differences. At this stage it should be considered whether the overall fold is good and if all areas can be modeled reliably. In some cases, certain areas lack a template and should be restricted into a specific secondary structure. An example of this can be seen in Toivola *et al.*, 2013, where a part of a long linker between two domains in the *Arabidopsis thaliana* NADPH-dependent thioredoxin reductase lacked a template structure and, hence, was restricted to an α-helix based on secondary structure predictions.

Kopp & Schwede (2004) and Mullins (2012) claim that the automated servers for protein modeling do not require as much expertise as non-automated modeling, which enables a broader audience to perform homology modeling. However, these types of servers are not to be trusted blindly. In the case of CIP2A (publication VI), I-TASSER is able to produce a model for the full-length sequence; however, when coupled to sequence analysis results, this model is not reliable. Protein modeling servers do facilitate the modeling work, especially when there is no detectable homolog with known structure, but the criteria and threshold values implemented in these servers should be carefully analyzed to see if they are met. Even more importantly, the researcher should know what each of these values describes and how reliable it is. Nevertheless, in the end, the model has to be carefully evaluated and examined both by evaluation programs and visually by the researcher.

5.9 Model quality

When assessing the model quality, multiple programs should be used to increase the reliability of the results. It is also important to know how the program or server works to be able to correctly interpret the relationship between the results and the target protein. Visual inspection of the model and comparison to the template are also vital procedures for adequate assessment of the model quality and possible improvements. The template structure should be carefully examined to unveil key amino acids, which stabilize the structure and, also, the residues contributing to the protein function. If these amino acids are conserved in the target protein, they are likely to have a

similar conformation as the corresponding amino acids in the template. This, in turn, makes the model better and increases the reliability of the results. For example, visual inspection of the CIP2A-ArmRP model (publication VI) revealed an unusual interaction pattern between an Arg residue and a hydrophobic environment. The conformation was corrected so that the Arg residue interacted with the solvent instead, which also enhanced the correctness of the electrostatic surface calculations. Furthermore, all published literature about experimental data on the target protein, the template structure and homologs should be carefully analyzed to help the modeling procedure and the quality and reliability assessment of the model. Important questions are whether the fold is logical and if the important amino acids are in the correct places. The coupling of the modeling results to the existing literature and experimental data cannot be done by computers, but rather have to be interpreted by the researcher to form conclusive results. Hence, knowledge and expertise about protein folding and the effects on function, especially when considering amino acid substitutions, are required. Furthermore, knowledge of structural biology in general is also important, as well as understanding of protein biochemistry, molecular interactions and specific biological phenomena related to each project.

The quality of structural models is often questioned, which is valid considering the number of errors that can be introduced in the model during each step of the process, especially if the sequence identity between the target and the template protein is low. However, structural genomics efforts focus on rapid structure determination to increase the number of protein families with a structurally characterized member in PDB (Mullins, 2012; Paliakasis *et al.*, 2008) and this has already resulted in an increase in the accuracy of homology models due to better structural homolog coverage. Furthermore, structural genomics have proven itself by structurally characterizing the majority of new families and contributing with five times as many novel folds as classical structural biology (Chandonia & Brenner, 2006; Gileadi *et al.*, 2007; Liu *et al.*, 2007; Marsden *et al.*, 2007; Todd *et al.*, 2005). The quality of the sequence alignment directly affects the quality and reliability of the structural model and, hence, sequence and structure alignment programs have become more sophisticated and still continue to improve, which in turn leads to more accurate protein structure predictions (Mullins, 2012). This is indeed needed, since the increase in new sequences provided by genome sequencing projects, coupled to the cost and tediousness of experimental structure solving, will make 3D structure modeling increasingly important as a tool for gaining insight into the structure of new proteins. Hence, there should be a careful selection of the targets in the structural genomics projects to ensure that the majority of the sequences could be modeled based on a template with at least 30 % sequence identity (Mullins, 2012).

There is also a need to improve the refinement process coupled to modeling. Now it is heavily dependent on the template protein, but it should rather consider the target structure to try to optimize the model as close as possible to the native state (Mullins, 2012). This could be beneficial for protein structures modeled with less than 30 % identity to the template, since these models may contain alignment errors, as well as have some real differences in structure compared to the template. Furthermore, the crystal structure is a snapshot of the protein in one state and may also contain errors in side chain conformations due to weak experimental data or flexibility. With this in mind, incorporating information implemented in electron density maps for template main chain when modeling proteins with low sequence identity to the template protein might improve the model. The maps could include the electron density for the side chains of conserved amino acids, as well as for amino acids of similar length, which would maximize the reliable information incorporated into the model. The implementation of electron density maps would also allow for identification of areas with weak electron density and lower confidence. These areas could then be less constrained and more flexible during the creation of the model.

Models are often considered to be highly hypothetical and not trustworthy before they have been verified by crystallization. Indeed, crystal structures may be important for verification of the modeling results when the overall quality is low, but other times the question at hand can be reliably answered with the help of models. In this work, CIP2A (publication VI) represents a study where a crystal structure or experimental data would be very beneficial to verify the results and get a more accurate and detailed knowledge about amino acids important for the function of the protein. This would markedly enhance the reliability for future drug development experiments. On the other hand, YeLpxR (publication III) represents a reliable model and the confidence of the results is high, which makes it unlikely that a crystallization experiment would add to the current knowledge. However, it cannot be ruled out that there could be some structural differences that are left undetected in the modeled structure. Hence, the need for validation of the modeling results by crystallization should be considered separately for each research project, especially when taking into account the time and money required for these experiments.

5.10 Inference of function from structural model

The function of an unknown protein is often inferred from evolutionary relationship to a homolog with characterized function. However, it should be kept in mind that protein families can be promiscuous: one fold might give multiple functions and several folds might exert one function (Todd *et al.*, 2001). Moreover, there are also examples of a similar structure, although there is no functional relationship or similarity between the protein sequences

(Sousounis *et al.*, 2012). The on-going structural genomics projects will certainly reveal more of these relationships that have been undetected at sequence level, since protein structure is much more conserved than sequence. However, functional variation has been shown to occur mostly when two proteins are less than 40 % identical. Also, even though the overall fold is highly conserved, substrate specificity and catalysis mechanisms might not be the same even if the sequence identity is higher than 50 % (Mullins, 2012; Rost, 2002; Tian & Skolnick, 2003), which is exemplified in the case of YeLpxR and StLpxR (publication III). Despite the high sequence identity (75 %), StLpxR can deacylate aminoarabinose-containing lipid A, while YeLpxR cannot. Hence, the experimental verification of the function inferred from structural modeling is still essential to distinguish the details behind the function of each protein. However, the computational function inference can serve as a valuable starting point for the experimental tests and a combination of the two methods enables maximum output and characterization.

Furthermore, bioinformatics enables researchers to design proteins with a desired function, which are then synthesized (Choong *et al.*, 2013). This touches upon the FucO project (publication I, II), which was aimed at designing an enzyme variant with activity towards the target substrate *S*-3-phenyl-1,2-propanediol. By redesigning the active site, the enzyme became a biocatalyst of the target substrate, although with moderate efficiency. The redesign was based on analysis of the available crystal structure, and then the active enzymes were modeled and subjected to docking studies. However, it is also possible to switch this workflow to analysis of the structure to find important amino acids, mutate these residues *in silico* and perform docking experiments to the mutants. The experimental work could then verify the results and build upon the best mutants obtained from the computational work. Hence, incorporating structural bioinformatics results can significantly reduce the experimental work and the related costs, since sequence alignments and homology models can pinpoint areas to be characterized experimentally. Moreover, corroborating experimental results always add to the reliability of the conclusions drawn from *in silico* studies and verify them. However, having experimental results at the beginning of the modeling process also benefits the modeling procedure. This way, the areas requiring extra attention are known and can be focused on when aligning sequences, thereby increasing the reliability of the models in the most crucial areas. Hence, many different factors and considerations should be taken into account in the modeling procedure, but with careful planning and execution the 3D structural models of proteins can greatly aid in obtaining important results.

6 Conclusions

In this thesis, five different proteins have been studied by 3D structural modeling: FucO, LpxR, LpxO, Slr0006 and CIP2A. They all represent separate case studies for different modeling techniques and highlight the possibilities for usage of a structural model.

In publication I and II, 3D structural models were created for FucO mutants and used for docking studies to explain differences in substrate specificities. The studies demonstrate the importance of additional space in the active site of FucO to install activity with *S*-3-phenyl-1,2-propanediol but, more importantly, they highlight the essentiality of retaining or maximizing the interactions between the enzyme and substrate. However, the role of the iron for the catalysis is still unknown and might have an effect on both the cofactor and the substrate binding and their positions. For future work, the catalytic mechanism should be studied in depth to verify the role of each individual component. Furthermore, from a computational perspective, the structure of FucO outside the active site could be studied more thoroughly and the amino acids predicted to affect the binding of *S*-3-phenyl-1,2-propanediol could be mutated *in silico* before performing docking studies to the mutants. The most promising mutants could then be tested experimentally to verify binding and catalysis.

In publication III, the homology model of YeLpxR was used for docking studies to explain the differences in substrate specificity compared to the ortholog StLpxR. The study pinpoints one amino acid, Asp31, as the main cause for the inability of YeLpxR to use aminoarabinose-containing lipid A for catalysis although StLpxR can use it. Asp31 limits the active site and makes it physically impossible for aminoarabinose-containing lipid A to fit into the active site pocket, thereby explaining the latency of YeLpxR at 21 °C. In this work, only the differing part of the lipid A molecules were docked to the protein and, therefore, it cannot be ruled out that other parts might affect the binding and positioning of the lipid A molecule in the binding site. Furthermore, we used rigid docking but including receptor flexibility in the docking protocol might give a slightly different and more exact positioning of the lipid A molecules in the active site. However, the aim of the project was to find out if lipid A decorated with aminoarabinose was physically unable to bind to YeLpxR and the reason for this. Hence, the docking method and settings were deemed to be good enough for this particular question. Furthermore, the results were ultimately verified by experimental studies and shown to be correct. Consequently, the modeled complex is not aimed to be detailed enough for studies on specific interactions and should not be used for such in depth analyses. Instead, the results give a general overview of the YeLpxR structure and where the amino

acids important for the substrate binding are located.

In publication IV, the model of LpxO was used to structurally and functionally characterize the protein. The fold was shown to be similar to human Asp/Asn β-hydroxylase, with an active site that resembles bovine Asp/Asn β-hydroxylase. Furthermore, catalytically important amino acids were mutated and experimentally proven to be essential. However, some of the mutated residues might be important for the correct positioning of one or more of the other catalytically important amino acids and not for the catalysis *per se*. This cannot be ruled out from the present results, but could be a valuable study in future work. Moreover, crystallization of LpxO would verify the structural model, as well as positioning of catalytically important amino acids.

In publication V, the homology model helped to determine that the Slr0006 protein belongs to the Sua5/YciO/YrdC family and has a similar active site to the YciO protein. Furthermore, the model showed that Slr0006 has a positively charged cleft, which possibly binds RNA or nucleotides. However, the aim was to solve the function and functional details of the Slr0006 protein, but these factors still remain uncertain. Therefore, future work should be concentrated to solving the function and performing crystallization studies to verify the structural details of Slr0006. Moreover, phylogenetics studies could help to determine the relationship between the Sua5/YciO/YrdC family members and Slr0006 and aid the correct classification of this protein.

In publication VI, the 3D structural model of CIP2A-ArmRP serves as a first insight into the structure-function relationship of the CIP2A protein. Existing literature was extensively incorporated into the analysis and evaluation of the model and showed that CIP2A-ArmRP is highly likely to follow the characteristics of other ArmRP proteins in forming interactions with partner proteins. The performed MD simulations might be regarded as too short to show without a doubt that the modeled CIP2A-ArmRP structure is stable, but coupled to the highly conserved and rigid structure of armadillo proteins in general, the performed parallel simulations were considered valuable enough. The obtained results are important for future work based upon structural information about CIP2A. Protein-peptide or protein-protein docking could be performed to analyze the binding mode of the identified interaction partners and experimental point mutations targeting the polar ladder could verify the binding site for some of the interaction partners. Furthermore, the model can help to create a stable construct for crystallization studies, which would be essential to verify the structural details of CIP2A-ArmRP.

The presented projects each represent a different modeling process based on the existence or nonexistence of a highly similar template. The projects

reflect the complexity of modeling, which comes from the fact that each project and protein presents its own challenges and cannot be studied in exactly the same manner. Often it is the method for obtaining a reliable sequence alignment that varies, which further supports its role as the most important step in the modeling process. However, caution and careful consideration should be a natural component in each of the steps in the modeling procedure (see Workflow 1). Of notice, it is important to incorporate as much information as possible from previously published literature or experimental data to maximize the knowledge gained from a protein structural model and to verify it. The lower the sequence identity, the more important it becomes to incorporate and interpret these factors within the modeling procedure to increase the reliability of the computational predictions. However, it is always important to understand and interpret the computational results yourself. Never leave it up to a computer because they cannot judge the data in the same way and many times they do not have the same understanding of a bigger picture. Conclusively, with knowledge on the performance of the computer programs, what they can do and their limitations, coupled to protein structure and chemistry knowledge, as well as critical analysis and interpretation of the results and their relationship to previously published data, structural bioinformatics can be the key to a successful project in a cost- and time-efficient manner.

MKYIILLIIVIAVLYVHYR
MFPELAILQQNWQVIRDEG
PSASQLCPQTTALLRDIPS
FIEV
RHI

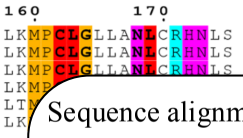
Sequence analysis

- Identify domains, signal peptides, membrane anchors etc.
- Secondary structure prediction
- Literature search

Query coverage	E value	Max ident
52%	5e-13	27%
28%		
31%		
14%		

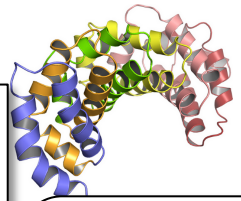
BLAST searches

- Sequences: start from smaller, manually annotated databases. Is the coverage good and E-value < 0.001?
- Crystal structures to be used as templates
  - Pay attention to sequence identity, coverage, resolution, ligands, cofactors, mutations, conformation, missing parts
- Does the secondary structure prediction match the template?
- Try multiple modeling webservers if no homologs with known structure are available



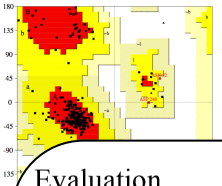
Sequence alignment

- Pairwise, multiple and/or structure-based
  - If multiple, include sequences from a range of 30-80 % sequence identity
- Edit to remove gaps in secondary structure elements
- Incorporate literature info: are functionally/structurally important amino acids conserved?



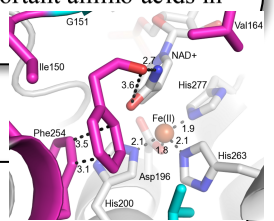
Create the model

- Create a set of 10 models (lowest energy usually best but check all)
- Considerations: is the overall fold good? Can all areas be modeled reliably? Should any areas be restricted?
- Models done with webservers: are the criteria and threshold values met? What does each value describe? How reliable is it?



Evaluation

- Use multiple programs to increase reliability. How does the program work? Is it suitable for your protein?
- Visually evaluate your model and compare to the template. Pay extra attention to amino acids of interest. Are conserved amino acids in the same conformation in template and model?
- Incorporate literature information. Is the fold logical? Are important amino acids in the right places?



Use of model

- Fold and charge distribution studies
- Docking studies
- Interpretation of experimental data
- Structure-function relationship analyses and explanations

References

Adamczak, R., Porollo, A., Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins,* **59**(3), 467-475.

Adams, R., Levine, I. (1923). Simplification of the gattermann synthesis of hydroxy aldehydes. *Journal of the American Chemical Society,* **45**(10), 2373-2377.

Agari, Y., Sato, S., Wakamatsu, T., Bessho, Y., Ebihara, A., Yokoyama, S., Kuramitsu, S., Shinkai, A. (2008). X-ray crystal structure of a hypothetical Sua5 protein from sulfolobus tokodaii strain 7. *Proteins,* **70**(3), 1108-1111.

Altman, R. B., Dugan, J. M. (2003). Defining bioinformatics and structural bioinformatics. *Methods of Biochemical Analysis,* **44**, 3-14.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology,* **215**(3), 403-410.

Altschul, S. F., Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends in Biochemical Sciences,* **23**(11), 444-447.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research,* **25**(17), 3389-3402.

Andrade, M. A., Petosa, C., O'Donoghue, S. I., Muller, C. W., Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *Journal of Molecular Biology,* **309**(1), 1-18.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., Murzin, A. G. (2014). SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Research,* **42**(Database issue), D310-4.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science,* **181**(4096), 223-230.

Apweiler, R., Bairoch, A., Wu, C. H. (2004). Protein sequence databases. *Current Opinion in Chemical Biology,* **8**(1), 76-80.

Armen, R. S., Chen, J., Brooks, C. L.,3rd. (2009). An evaluation of explicit receptor flexibility in molecular docking using molecular dynamics and torsion angle molecular dynamics. *Journal of Chemical Theory and Computation,* **5**(10), 2909-2923.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., Notredame, C. (2006). Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-coffee. *Nucleic Acids Research,* **34**(Web Server issue), W604-8.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N. (2010). ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research,* **38**(Web Server issue), W529-33.

Aussel, L., Therisod, H., Karibian, D., Perry, M. B., Bruneteau, M., Caroff, M. (2000). Novel variation of lipid A structures in strains of different yersinia species. *FEBS Letters,* **465**(1), 87-92.

Azam, S. S., Abbasi, S. W. (2013). Molecular docking studies for the identification of novel melatoninergic inhibitors for acetylserotonin-O-methyltransferase using different docking routines. *Theoretical Biology & Medical Modelling,* **10**, 63-4682-10-63.

Baldoma, L., Aguilar, J. (1988). Metabolism of L-fucose and L-rhamnose in escherichia coli: Aerobic-anaerobic regulation of L-lactaldehyde dissimilation. *Journal of Bacteriology,* **170**(1), 416-421.

Barre-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., *et al*. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science,* **220**(4599), 868-871.

Battchikova, N., Vainonen, J. P., Vorontsova, N., Keranen, M., Carmel, D., Aro, E. M. (2010). Dynamic changes in the proteome of synechocystis 6803 in response to CO(2) limitation revealed by quantitative proteomics. *Journal of Proteome Research,* **9**(11), 5896-5912.

Baxevanis, A. D., & Ouellette, B. F. F. (2004). *Bioinformatics: A practical guide to the analysis of genes and proteins* (3rd ed.). Hoboken N.J.: Wiley-Interscience.

Bengoechea, J. A., Brandenburg, K., Arraiza, M. D., Seydel, U., Skurnik, M., Moriyon, I. (2003). Pathogenic yersinia enterocolitica strains increase the outer membrane permeability in response to environmental stimuli by modulating lipopolysaccharide fluidity and lipid A structure. *Infection and Immunity,* **71**(4), 2014-2021.

Benkert, P., Kunzli, M., Schwede, T. (2009). QMEAN server for protein model quality estimation. *Nucleic Acids Research,* **37**(Web Server issue), W510-4.

Berendsen, H., Postma, J., van Gunsteren, W., DiNola, A., Haak, J. (1984). Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics,* **81**, 3684.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S.*, et al*. (2002). The protein data bank. *Acta Crystallographica Section D, Biological Crystallography,* **58**(Pt 6 No 1), 899-907.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., Westbrook, J. (2000). The protein data bank and the challenge of structural genomics. *Nature Structural Biology,* **7 Suppl**, 957-959.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F.,Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977). The protein data bank. A computer-based archival file for macromolecular structures. *European Journal of Biochemistry / FEBS,* **80**(2), 319-324.

Biegert, A., Soding, J. (2009). Sequence context-specific profiles for homology searching. *Proceedings of the National Academy of Sciences of the United States of America,* **106**(10), 3770-3775.

Bissantz, C., Schalon, C., Guba, W., Stahl, M. (2005). Focused library design in GPCR projects on the example of 5-HT(2c) agonists: Comparison of structure-based virtual screening with ligand-based search methods. *Proteins,* **61**(4), 938-952.

Blank, L. M., Ebert, B. E., Buehler, K., Buhler, B. (2010). Redox biocatalysis and metabolism: Molecular mechanisms and metabolic network analysis. *Antioxidants & Redox Signaling,* **13**(3), 349-394.

Blikstad, C., Widersten, M. (2010). Functional characterization of a stereospecific diol dehydrogenase, FucO, from *Escherichia coli*: substrate specificity, pH dependence, kinetic isotope effects and influence of solvent viscosity. *Journal of Molecular Catalysis B: Enzymatic,* **66**, 148-155.

Blikstad, C., Dahlström, K. M., Salminen, T. A., Widersten, M. (2014). Substrate scope and selectivity in offspring to an enzyme subjected to directed evolution. *The FEBS Journal,* **281**(10), 2387-2398.

Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., Sutcliffe, M. (1988). 18th sir hans krebs lecture. knowledge-based protein modelling and design. *European Journal of Biochemistry / FEBS,* **172**(3), 513-520.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J., Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature,* **326**(6111), 347-352.

Bockelman, C., Lassus, H., Hemmes, A., Leminen, A., Westermarck, J., Haglund, C., Butzow, R., Ristimaki, A. (2011). Prognostic role of CIP2A expression in serous ovarian cancer. *British Journal of Cancer,* **105**(7), 989-995.

Böhm, H. J., & Stahl, M. (2002). The use of scoring functions in drug discovery applications. In K. B. Lipkowitz, & D. B. Boyd (Eds.), *Reviews in computational chemistry* (Vol 18 ed., pp. 41-88). New Jersey: Wiley-VCH Inc.

Bonneau, R., Baker, D. (2001). Ab initio protein structure prediction: Progress and prospects. *Annual Review of Biophysics and Biomolecular Structure,* **30**, 173-189.

Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., Robins, K. (2012). Engineering the third wave of biocatalysis. *Nature,* **485**(7397), 185-194.

Boronat, A., Aguilar, J. (1979). Rhamnose-induced propanediol oxidoreductase in escherichia coli: Purification, properties, and comparison with the fucose-induced enzyme. *Journal of Bacteriology,* **140**(2), 320-326.

Bottone, E. J. (1997). Yersinia enterocolitica: The charisma continues. *Clinical Microbiology Reviews,* **10**(2), 257-276.

Bowie, J. U., Luthy, R., Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**(5016), 164-170.

Buenavista, M. T., Roche, D. B., McGuffin, L. J. (2012). Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. *Bioinformatics,* **28**(14), 1851-1857.

Callaway, E. (2015). The revolution will not be crystallized: A new method sweeps through structural biology. *Nature,* **525**(7568), 172-174.

Carlsson, J., Coleman, R. G., Setola, V., Irwin, J. J., Fan, H., Schlessinger, A., Sali, A., Roth, B. L., Shoichet, B. K. (2011). Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nature Chemical Biology,* **7**(11), 769-778.

Carmel, D., Battchikova, N., Holmström, M., Mulo, P., & Aro, E. M. (2012). Knock-out of low CO_2-induced *slr0006* gene in *Synechocystis* sp. PCC 6803: Consequences on growth and proteome. In C. Lu (Ed.), *Photosynthesis: Research for food, fuel and future - 15th international conference on photosynthesis.* New York: Zhejiang University Press, Springer GmbH.

Carmel, D., Mulo, P., Battchikova, N., Aro, E. M. (2011). Membrane attachment of Slr0006 in synechocystis sp. PCC 6803 is determined by divalent ions. *Photosynthesis Research,* **108**(2-3), 241-245.

Carpenter, B., Hemsworth, G. R., Wu, Z., Maamra, M., Strasburger, C. J., Ross, R. J., Artymiuk, P. J. (2012). Structure of the human obesity receptor leptin-binding domain reveals the mechanism of leptin antagonism by a monoclonal antibody. *Structure,* **20**(3), 487-497.

Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M.*, et al*. (2012). AMBER12. *University of California, San Fransisco,*

Celniker, G. (2013). ConSurf: Using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry,* **53**(3), 199-206.

Chan, H. S., Dill, K. A. (1990). Origins of structure in globular proteins. *Proceedings of the National Academy of Sciences of the United States of America,* **87**(16), 6388-6392.

Chandonia, J. M., Brenner, S. E. (2006). The impact of structural genomics: Expectations and outcomes. *Science,* **311**(5759), 347-351.

Chang, K. W., Tsai, T. Y., Chen, K. C., Yang, S. C., Huang, H. J., Chang, T. T., Sun, M. F., Chen, H. Y., Tsai, F. J., Chen, C. Y. (2011). iSMART: An integrated cloud computing web server for traditional chinese medicine for online virtual screening, de novo evolution and drug design. *Journal of Biomolecular Structure & Dynamics,* **29**(1), 243-250.

Chen, H., Kihara, D. (2008). Estimating quality of template-based protein models by alignment stability. *Proteins,* **71**(3), 1255-1274.

Chen, H. Y., Chang, S. S., Chan, Y. C., Chen, C. Y. (2014). Discovery of novel insomnia leads from screening traditional chinese medicine database. *Journal of Biomolecular Structure & Dynamics,* **32**(5), 776-791.

Chen, Y. C. (2015). Beware of docking! *Trends in Pharmacological Sciences,* **36**(2), 78-95.

Chen, Z., Zhao, H. (2005). Rapid creation of a novel protein function by in vitro coevolution. *Journal of Molecular Biology,* **348**(5), 1273-1282.

Cheng, J., Wang, Z., Tegge, A. N., Eickholt, J. (2009). Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins,* **77 Suppl 9**, 181-184.

Cheng, T., Li, X., Li, Y., Liu, Z., Wang, R. (2009). Comparative assessment of scoring functions on a diverse test set. *Journal of Chemical Information and Modeling,* **49**(4), 1079-1093.

Choong, Y. S., Tye, G. J., Lim, T. S. (2013). Minireview: Applied structural bioinformatics in proteomics. *The Protein Journal,* **32**(7), 505-511.

Chothia, C., Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal,* **5**(4), 823-826.

Chou, P. Y., Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry,* **13**(2), 222-245.

Claessens, M., Van Cutsem, E., Lasters, I., Wodak, S. (1989). Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Engineering,* **2**(5), 335-345.

Clements, A., Tull, D., Jenney, A. W., Farn, J. L., Kim, S. H., Bishop, R. E., McPhee, J. B., Hancock, R. E., Hartland, E. L., Pearse, M. J.*, et al*. (2007). Secondary acylation of klebsiella pneumoniae lipopolysaccharide contributes to sensitivity to antibacterial peptides. *The Journal of Biological Chemistry,* **282**(21), 15569-15577.

Coates, J. C. (2003). Armadillo repeat proteins: Beyond the animal kingdom. *Trends in Cell Biology,* **13**(9), 463-471.

Coffin, J., Haase, A., Levy, J. A., Montagnier, L., Oroszlan, S., Teich, N., Temin, H., Toyoshima, K., Varmus, H., Vogt, P. (1986). Human immunodeficiency viruses. *Science,* **232**(4751), 697.

Cole, C., Barber, J. D., Barton, G. J. (2008). The jpred 3 secondary structure prediction server. *Nucleic Acids Research,* **36**(Web Server issue), W197-201.

Colovos, C., Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science,* **2**(9), 1511-1519.

Come, C., Laine, A., Chanrion, M., Edgren, H., Mattila, E., Liu, X., Jonkers, J., Ivaska, J., Isola, J., Darbon, J. M.*, et al*. (2009). CIP2A is associated with human breast cancer aggressivity. *Clinical Cancer Research,* **15**(16), 5092-5100.

Conti, E., Uy, M., Leighton, L., Blobel, G., Kuriyan, J. (1998). Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha. *Cell,* **94**(2), 193-204.

Conway, T., Ingram, L. O. (1989). Similarity of escherichia coli propanediol oxidoreductase (fucO product) and an unusual alcohol dehydrogenase

from zymomonas mobilis and saccharomyces cerevisiae. *Journal of Bacteriology,* **171**(7), 3754-3759.

Counterman, A. E., Clemmer, D. E. (1999). Volumes of individual amino acid residues in gas-phase peptide ions. *Journal of the American Chemical Society,* **121**, 4031-4039.

Creighton, T. E. (1990). Protein folding. *The Biochemical Journal,* **270**(1), 1-16.

Cutress, M. L., Whitaker, H. C., Mills, I. G., Stewart, M., Neal, D. E. (2008). Structural basis for the nuclear import of the human androgen receptor. *Journal of Cell Science,* **121**(Pt 7), 957-968.

Damm-Ganamet, K. L., Smith, R. D., Dunbar, J. B.,Jr, Stuckey, J. A., Carlson, H. A. (2013). CSAR benchmark exercise 2011-2012: Evaluation of results from docking and relative ranking of blinded congeneric series. *Journal of Chemical Information and Modeling,* **53**(8), 1853-1870.

Daugelaite, J., O'Driscoll, A., Sleator, R. D. (2013). An overview of multiple sequence alignments and cloud computing in bioinformatics. *ISRN Biomathematics,* **2013**(Article ID 615630)

Dayhoff, M. (1978). *Atlas of protein sequence and structure, volume 5*. Washington (D.C.): National Biomedical Research Foundation.

Dayhoff, M., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure* (pp. 345–352). Washington, DC: National Biomedical Research Foundation.

Dayhoff, M., Eck, R. (1968). *Atlas of protein sequence and structure 1967-1968*. Maryland (Silver Spring): National Biomedical Research Foundation.

de Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., Hulo, N. (2006). ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research,* **34**(Web Server issue), W362-5.

de Graaf, C., Foata, N., Engkvist, O., Rognan, D. (2008). Molecular modeling of the second extracellular loop of G-protein coupled receptors and its implication on structure-based virtual screening. *Proteins,* **71**(2), 599-620.

De Majumdar, S., Yu, J., Fookes, M., McAteer, S. P., Llobet, E., Finn, S., Spence, S., Monahan, A., Kissenpfennig, A., Ingram, R. J.*, et al*. (2015).

Elucidation of the RamA regulon in klebsiella pneumoniae reveals a role in LPS regulation. *PLoS Pathogens,* **11**(1), e1004627.

De, P., Carlson, J., Leyland-Jones, B., Dey, N. (2014). Oncogenic nexus of cancerous inhibitor of protein phosphatase 2A (CIP2A): An oncoprotein with many hands. *Oncotarget,* **5**(13), 4581-4602.

di Luccio, E., Koehl, P. (2011). A quality metric for homology modeling: The H-factor. *BMC Bioinformatics,* **12**, 48-2105-12-48.

Dolan, M. A., Noah, J. W., Hurt, D. (2012). Comparison of common homology modeling algorithms: Application of user-defined alignments. *Methods in Molecular Biology,* **857**, 399-414.

Dong, Q. Z., Wang, Y., Dong, X. J., Li, Z. X., Tang, Z. P., Cui, Q. Z., Wang, E. H. (2011). CIP2A is overexpressed in non-small cell lung cancer and correlates with poor prognosis. *Annals of Surgical Oncology,* **18**(3), 857-865.

Dorn, M., E Silva, M. B., Buriol, L. S., Lamb, L. C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. *Computational Biology and Chemistry,* **53PB**, 251-276.

Duan, Y., Wu, C., Chowdhury, S., Lee, M., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T.*, et al*. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mehcanical calculations. *Journal of Computational Chemistry,* **24**, 1999.

Dunbar, J. B.,Jr, Smith, R. D., Damm-Ganamet, K. L., Ahmed, A., Esposito, E. X., Delproposto, J., Chinnaswamy, K., Kang, Y. N., Kubish, G., Gestwicki, J. E.*, et al*. (2013). CSAR data set release 2012: Ligands, affinities, complexes, and docking decoys. *Journal of Chemical Information and Modeling,* **53**(8), 1842-1852.

Durrant, J. D., McCammon, J. A. (2010). NNScore: A neural-network-based scoring function for the characterization of protein-ligand complexes. *Journal of Chemical Information and Modeling,* **50**(10), 1865-1871.

Durrant, J. D., McCammon, J. A. (2011). NNScore 2.0: A neural-network receptor-ligand scoring function. *Journal of Chemical Information and Modeling,* **51**(11), 2897-2903.

Edman, P., Begg, G. (1967). A protein sequenator. *European Journal of Biochemistry / FEBS,* **1**(1), 80-91.

Eisenberg, D., Luthy, R., Bowie, J. U. (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods in Enzymology,* **277**, 396-404.

Eklof Spink, K., Fridman, S. G., Weis, W. I. (2001). Molecular mechanisms of beta-catenin recognition by adenomatous polyposis coli revealed by the structure of an APC-beta-catenin complex. *The EMBO Journal,* **20**(22), 6203-6212.

El Yacoubi, B., Hatin, I., Deutsch, C., Kahveci, T., Rousset, J. P., Iwata-Reuyl, D., Murzin, A. G., de Crecy-Lagard, V. (2011). A role for the universal Kae1/Qri7/YgjD (COG0533) family in tRNA modification. *The EMBO Journal,* **30**(5), 882-893.

El Yacoubi, B., Lyons, B., Cruz, Y., Reddy, R., Nordin, B., Agnelli, F., Williamson, J. R., Schimmel, P., Swairjo, M. A., de Crecy-Lagard, V. (2009). The universal YrdC/Sua5 family is required for the formation of threonylcarbamoyladenosine in tRNA. *Nucleic Acids Research,* **37**(9), 2894-2909.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., Mee, R. P. (1997). Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design,* **11**(5), 425-445.

Enders, D., Bhushan, V. (1988). Enantioselective synthesis of protected alpha-hydroxy aldehydes and ketones via hydroxylation of metalated chiral hydrazones. *Tetrahedron Letters,* **29**(20), 2437-2440.

Engel, S., Skoumbourdis, A. P., Childress, J., Neumann, S., Deschamps, J. R., Thomas, C. J., Colson, A. O., Costanzi, S., Gershengorn, M. C. (2008). A virtual screen for diverse ligands: Discovery of selective G protein-coupled receptor antagonists. *Journal of the American Chemical Society,* **130**(15), 5115-5123.

Essmann, U., Perera, L., Berkowitz, M., Darden, T., Lee, H., Pedersen, L. (1995). A smooth particle mesh ewald method. *Journal of Chemical Physics,* **103**, 8577.

Feng, D. F., Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution,* **25**(4), 351-360.

Ferguson, A. D., Hofmann, E., Coulton, J. W., Diederichs, K., Welte, W. (1998). Siderophore-mediated iron transport: Crystal structure of FhuA with bound lipopolysaccharide. *Science,* **282**(5397), 2215-2220.

Ferreira, L. G., Dos Santos, R. N., Oliva, G., Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. *Molecules,* **20**(7), 13384-13421.

Finkelstein, A. V., Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Progress in Biophysics and Molecular Biology,* **50**(3), 171-190.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J.*, et al*. (2014). Pfam: The protein families database. *Nucleic Acids Research,* **42**(Database issue), D222-30.

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

Flohil, J. A., Vriend, G., Berendsen, H. J. (2002). 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. *Proteins,* **48**(4), 593-604.

Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M., Rajgaria, R. (2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science,* **61**, 966 – 988.

Fredman, M. L. (1984). Algorithms for computing evolutionary similarity measures with length independent gap penalties. *Bulletin of Mathematical Biology,* **46**(4), 553-566.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K.*, et al*. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry,* **47**(7), 1739-1749.

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., Mainz, D. T. (2006). Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry,* **49**(21), 6177-6196.

Fu, T. M., Liu, X., Li, L., Su, X. D. (2010). The structure of the hypothetical protein smu.1377c from streptococcus mutans suggests a role in tRNA modification. *Acta Crystallographica.Section F, Structural Biology and Crystallization Communications,* **66**(Pt 7), 771-775.

Gallo, R. C., Montagnier, L. (1988). AIDS in 1988. *Scientific American,* **259**(4), 41-48.

Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G. M., Kaplan, M., Haynes, B. F., Palker, T. J., Redfield, R., Oleske, J., Safai, B. (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III)

from patients with AIDS and at risk for AIDS. *Science,* **224**(4648), 500-503.

Galperin, M. Y., Koonin, E. V. (2004). 'Conserved hypothetical' proteins: Prioritization of targets for experimental study. *Nucleic Acids Research,* **32**(18), 5452-5463.

Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J., Freer, S. T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chemistry & Biology,* **2**(5), 317-324.

Gibas, C., & Jambeck, P. (2001). *Developing bioinformatics computer skills*. Sebastopol CA, O'Reilly.

Gibbons, H. S., Lin, S., Cotter, R. J., Raetz, C. R. (2000). Oxygen requirement for the biosynthesis of the S-2-hydroxymyristate moiety in salmonella typhimurium lipid A. function of LpxO, A new Fe2+/alpha-ketoglutarate-dependent dioxygenase homologue. *The Journal of Biological Chemistry,* **275**(42), 32940-32949.

Gibbons, H. S., Reynolds, C. M., Guan, Z., Raetz, C. R. (2008). An inner membrane dioxygenase that generates the 2-hydroxymyristate moiety of salmonella lipid A. *Biochemistry,* **47**(9), 2814-2825.

Gileadi, O., Knapp, S., Lee, W. H., Marsden, B. D., Muller, S., Niesen, F. H., Kavanagh, K. L., Ball, L. J., von Delft, F., Doyle, D. A.*, et al*. (2007). The scientific impact of the structural genomics consortium: A protein family and ligand-centered approach to medically-relevant human proteins. *Journal of Structural and Functional Genomics,* **8**(2-3), 107-119.

Giribet, G., Wheeler, W. C. (1999). On gaps. *Molecular Phylogenetics and Evolution,* **13**(1), 132-143.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., Ben-Tal, N. (2003). ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics,* **19**(1), 163-164.

Gohlke, H., Hendlich, M., Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology,* **295**(2), 337-356.

Goldberg, K., Schroer, K., Lutz, S., Liese, A. (2007). Biocatalytic ketone reduction--a powerful tool for the production of chiral alcohols--part I: Processes with isolated enzymes. *Applied Microbiology and Biotechnology,* **76**(2), 237-248.

Goodsell, D. S., Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins,* **8**(3), 195-202.

Gouet, P., Courcelle, E., Stuart, D. I., Metoz, F. (1999). ESPript: Analysis of multiple sequence alignments in PostScript. *Bioinformatics,* **15**(4), 305-308.

Graham, T. A., Ferkey, D. M., Mao, F., Kimelman, D., Xu, W. (2001). Tcf4 can specifically recognize beta-catenin using alternative conformations. *Nature Structural Biology,* **8**(12), 1048-1052.

Graham, T. A., Weaver, C., Mao, F., Kimelman, D., Xu, W. (2000). Crystal structure of a beta-catenin/tcf complex. *Cell,* **103**(6), 885-896.

Greer, J. (1990). Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins,* **7**(4), 317-334.

Gribskov, M., McLachlan, A. D., Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America,* **84**(13), 4355-4358.

Guedes, I. A., de Magalhães, C. S., Dardenne, L. E. (2014). Receptor-ligand molecular docking. *Biophysical Reviews,* **6**, 75-87.

Guex, N., Peitsch, M. C. (1997). SWISS-MODEL and the swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis,* **18**(15), 2714-2723.

Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews Genetics,* **1**(3), 231-236.

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., Banks, J. L. (2004). Glide: A new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of Medicinal Chemistry,* **47**(7), 1750-1759.

Hall, M., Bommarius, A. S. (2011). Enantioenriched compounds via enzyme-catalyzed redox reactions. *Chemical Reviews,* **111**(7), 4088-4110.

Hanson, M. A., Stevens, R. C. (2009). Discovery of new GPCR biology: One receptor structure at a time. *Structure,* **17**(1), 8-14.

Hayat, S., Elofsson, A. (2012). BOCTOPUS: Improved topology prediction of transmembrane beta barrel proteins. *Bioinformatics,* **28**(4), 516-522.

He, H., Wu, G., Li, W., Cao, Y., Liu, Y. (2012). CIP2A is highly expressed in hepatocellular carcinoma and predicts poor prognosis. *Diagnostic Molecular Pathology : The American Journal of Surgical Pathology, Part B,* **21**(3), 143-149.

Henikoff, S., Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America,* **89**(22), 10915-10919.

Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., Raushel, F. M. (2007). Structure-based activity prediction for an enzyme of unknown function. *Nature,* **448**(7155), 775-779.

Hillisch, A., Pineda, L. F., Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. *Drug Discovery Today,* **9**(15), 659-669.

Hofmann, K., Stoffel, W. (1993). TMbase - A database of membrane spanning proteins segments. *Biological Chemistry,* **374**, 166.

Hofmann, K. (2000). Sensitive protein comparisons with profiles and hidden markov models. *Briefings in Bioinformatics,* **1**(2), 167-178.

Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Computational Biology,* **7**(3), e1002021.

Holland, T. A., Veretnik, S., Shindyalov, I. N., Bourne, P. E. (2006). Partitioning protein structures into domains: Why is it so difficult? *Journal of Molecular Biology,* **361**(3), 562-590.

Holm, L., Sanders, C. (1996). Mapping the protein universe. *Nature,* **273**(5275), 595-603.

Hooft, R. W., Vriend, G., Sander, C., Abola, E. E. (1996). Errors in protein structures. *Nature,* **381**(6580), 272.

Hoyos, P., Sinisterra, J. V., Molinari, F., Alcantara, A. R., Dominguez de Maria, P. (2010). Biocatalytic strategies for the asymmetric synthesis of alpha-hydroxy ketones. *Accounts of Chemical Research,* **43**(2), 288-299.

Huang, B. (2009). MetaPocket: A meta approach to improve protein ligand binding site prediction. *Omics : A Journal of Integrative Biology,* **13**(4), 325-330.

Huang, Y. J., Mao, B., Aramini, J. M., Montelione, G. T. (2014). Assessment of template-based protein structure predictions in CASP10. *Proteins,* **82 Suppl 2**, 43-56.

Hubbard, T. J., Lesk, A. M., Tramontano, A. (1996). Gathering them in to the fold. *Nature Structural Biology,* **3**(4), 313.

Huff, J. R., Kahn, J. (2001). Discovery and clinical development of HIV-1 protease inhibitors. *Advances in Protein Chemistry,* **56**, 213-251.

Humphrey, W., Dalke, A., Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics,* **14**(1), 33-8, 27-8.

Illergard, K., Ardell, D. H., Elofsson, A. (2009). Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins,* **77**(3), 499-508.

Ingalls, A. M., Dickie, M. M., Snell, G. D. (1950). Obese, a new mutation in the house mouse. *The Journal of Heredity,* **41**(12), 317-318.

Iserentant, H., Peelman, F., Defeau, D., Vandekerckhove, J., Zabeau, L., Tavernier, J. (2005). Mapping of the interface between leptin and the leptin receptor CRH2 domain. *Journal of Cell Science,* **118**(Pt 11), 2519-2527.

Ishiyama, N., Lee, S. H., Liu, S., Li, G. Y., Smith, M. J., Reichardt, L. F., Ikura, M. (2010). Dynamic and static interactions between p120 catenin and E-cadherin regulate the stability of cell-cell adhesion. *Cell,* **141**(1), 117-128.

Jia, J., Lunin, V. V., Sauve, V., Huang, L. W., Matte, A., Cygler, M. (2002). Crystal structure of the YciO protein from escherichia coli. *Proteins,* **49**(1), 139-141.

Jiang, F., Kim, S. H. (1991). "Soft docking": Matching of molecular surface cubes. *Journal of Molecular Biology,* **219**(1), 79-102.

Johnson, M. S., & Lehtonen, J. V. (2000). Comparison of protein three-dimensional structures. In D. Higgins, & W. Taylor (Eds.), *Bioinformatics: Sequence, structure and databanks.* (pp. 15). Oxford, UK: Oxford University Press.

Johnson, M. S., May, A. C., Rodionov, M. A., Overington, J. P. (1996). Discrimination of common protein folds: Application of protein structure to sequence/structure comparisons. *Methods in Enzymology,* **266**, 575-598.

Johnson, M. S., Overington, J. P. (1993). A structural basis for sequence comparisons. an evaluation of scoring methodologies. *Journal of Molecular Biology,* **233**(4), 716-738.

Johnson, M. S., Overington, J. P., Blundell, T. L. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *Journal of Molecular Biology,* **231**(3), 735-752.

Johnson, M. S., Srinivasan, N., Sowdhamini, R., Blundell, T. L. (1994). Knowledge-based protein modeling. *Critical Reviews in Biochemistry and Molecular Biology,* **29**(1), 1-68.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology,* **292**(2), 195-202.

Jones, G., Willett, P., Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology,* **245**(1), 43-53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology,* **267**(3), 727-748.

Jones, T. A., Thirup, S. (1986). Using known substructures in protein model building and crystallography. *The EMBO Journal,* **5**(4), 819-822.

Jorgensen, W., Chandrasekhar, J., Madura, J., Impey, R., Klein, M. (1983). Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics,* **79**, 926.

Junttila, M. R., Puustinen, P., Niemela, M., Ahola, R., Arnold, H., Bottzauw, T., Ala-aho, R., Nielsen, C., Ivaska, J., Taya, Y.*, et al*. (2007). CIP2A inhibits PP2A in human malignancies. *Cell,* **130**(1), 51-62.

Kalman, M., Ben-Tal, N. (2010). Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics,* **26**(10), 1299-1307.

Kapetanovic, I. M. (2008). Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chemico-Biological Interactions,* **171**(2), 165-176.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry,* **14**, 1-63.

Kawahara, K., Tsukano, H., Watanabe, H., Lindner, B., Matsuura, M. (2002). Modification of the structure and activity of lipid A in yersinia pestis lipopolysaccharide by growth temperature. *Infection and Immunity,* **70**(8), 4092-4098.

Kawasaki, K., Ernst, R. K., Miller, S. I. (2005). Inhibition of salmonella enterica serovar typhimurium lipopolysaccharide deacylation by aminoarabinose membrane modification. *Journal of Bacteriology,* **187**(7), 2448-2457.

Kawasaki, K., Teramoto, M., Tatsui, R., Amamoto, S. (2012). Lipid A 3'-O-deacylation by salmonella outer membrane enzyme LpxR modulates the ability of lipid A to stimulate toll-like receptor 4. *Biochemical and Biophysical Research Communications,* **428**(3), 343-347.

Kellenberger, E., Springael, J. Y., Parmentier, M., Hachet-Haas, M., Galzi, J. L., Rognan, D. (2007). Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *Journal of Medicinal Chemistry,* **50**(6), 1294-1303.

Kelley, L. A., Sternberg, M. J. (2009). Protein structure prediction on the web: A case study using the phyre server. *Nature Protocols,* **4**(3), 363-371.

Khanna, A., Bockelman, C., Hemmes, A., Junttila, M. R., Wiksten, J. P., Lundin, M., Junnila, S., Murphy, D. J., Evan, G. I., Haglund, C.*, et al*. (2009). MYC-dependent regulation and prognostic role of CIP2A in gastric cancer. *Journal of the National Cancer Institute,* **101**(11), 793-805.

Khanna, A., Okkeri, J., Bilgen, T., Tiirikka, T., Vihinen, M., Visakorpi, T., Westermarck, J. (2011). ETS1 mediates MEK1/2-dependent overexpression of cancerous inhibitor of protein phosphatase 2A (CIP2A) in human cancer cells. *PloS One,* **6**(3), e17979.

Khanna, A., Pimanda, J. E., Westermarck, J. (2013). Cancerous inhibitor of protein phosphatase 2A, an emerging human oncoprotein and a potential cancer therapy target. *Cancer Research,* **73**(22), 6548-6553.

Kim, D. E., Chivian, D., Baker, D. (2004). Protein structure prediction and analysis using the robetta server. *Nucleic Acids Research,* **32**(Web Server issue), W526-31.

Kopp, J., Bordoli, L., Battey, J. N., Kiefer, F., Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins,* **69 Suppl 8**, 38-56.

Kopp, J., Schwede, T. (2004). Automated protein structure homology modeling: A progress report. *Pharmacogenomics,* **5**(4), 405-416.

Korb, O., Stutzle, T., Exner, T. E. (2009). Empirical scoring functions for advanced protein-ligand docking with PLANTS. *Journal of Chemical Information and Modeling,* **49**(1), 84-96.

Koshland, D. E.,Jr, Nemethy, G., Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry,* **5**(1), 365-385.

Kratochwil, N. A., Malherbe, P., Lindemann, L., Ebeling, M., Hoener, M. C., Muhlemann, A., Porter, R. H., Stahl, M., Gerber, P. R. (2005). An automated system for the analysis of G protein-coupled receptor transmembrane binding pockets: Alignment, receptor-based pharmacophores, and their application. *Journal of Chemical Information and Modeling,* **45**(5), 1324-1336.

Kratzer, R., Nidetzky, B. (2007). Identification of candida tenuis xylose reductase as highly selective biocatalyst for the synthesis of aromatic alpha-hydroxy esters and improvement of its efficiency by protein engineering. *Chemical Communications,* **10**, 1047-1049.

Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T., Tramontano, A. (2015). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins,* doi: 10.1002/prot.24919.

Kuhlbrandt, W. (2013). Introduction to electron crystallography. *Methods in Molecular Biology,* **955**, 1-16.

Kumar, S., Ma, B., Tsai, C. J., Sinha, N., Nussinov, R. (2000). Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Science,* **9**(1), 10-19.

Kuratani, M., Kasai, T., Akasaka, R., Higashijima, K., Terada, T., Kigawa, T., Shinkai, A., Bessho, Y., Yokoyama, S. (2011). Crystal structure of sulfolobus tokodaii Sua5 complexed with L-threonine and AMPPNP. *Proteins,* **79**(7), 2065-2075.

Kurczab, R., Nowak, M., Chilmonczyk, Z., Sylte, I., Bojarski, A. J. (2010). The development and validation of a novel virtual screening cascade protocol to identify potential serotonin 5-HT(7)R antagonists. *Bioorganic & Medicinal Chemistry Letters,* **20**(8), 2465-2468.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., Ben-Tal, N. (2005). ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research,* **33**(Web Server issue), W299-302.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W*., et al*. (2001). Initial sequencing and analysis of the human genome. *Nature,* **409**(6822), 860-921.

Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J. R., Whittle, P. J., Danley, D. E., Geoghegan, K. F. (1989). X-ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes. *Nature,* **342**(6247), 299-302.

Lape, M., Elam, C., Paula, S. (2010). Comparison of current docking tools for the simulation of inhibitor binding by the transmembrane domain of the sarco/endoplasmic reticulum calcium ATPase. *Biophysical Chemistry,* **150**(1-3), 88-97.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R*., et al*. (2007). Clustal W and clustal X version 2.0. *Bioinformatics,* **23**(21), 2947-2948.

Laskowski, R. A. (1995). SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics,* **13**(5), 323-30, 307-8.

Laskowski, R. A., Moss, D. S., Thornton, J. M. (1993). Main-chain bond lengths and bond angles in protein structures. *Journal of Molecular Biology,* **231**(4), 1049-1067.

Lassmann, T., Sonnhammer, E. L. (2005). Automatic assessment of alignment quality. *Nucleic Acids Research,* **33**(22), 7120-7128.

Laurents, D. V., Subbiah, S., Levitt, M. (1994). Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein Science,* **3**(11), 1938-1944.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *Journal of Molecular Biology,* **235**(1), 345-356.

Lehtonen, J. V., Still, D. J., Rantanen, V. V., Ekholm, J., Bjorklund, D., Iftikhar, Z., Huhtala, M., Repo, S., Jussila, A., Jaakkola, J.*, et al*. (2004). BODIL: A molecular modeling environment for structure-function analysis and drug design. *Journal of Computer-Aided Molecular Design,* **18**(6), 401-419.

Lesk, A. M. (2000). *Introduction to protein architecture: The structural biology of proteins*. Oxford, Oxford University Press.

Lesk, A. M. (2002). *Introduction to bioinformatics*. Oxford, Oxford University Press.

Letunic, I., Doerks, T., Bork, P. (2015). SMART: Recent updates, new developments and status in 2015. *Nucleic Acids Research,* **43**(Database issue), D257-60.

Levinthal, C. (1966). Molecular model-building by computer. *Scientific American,* **214**(6), 42-52.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology,* **226**(2), 507-533.

Levitt, M., Chothia, C. (1976). Structural patterns in globular proteins. *Nature,* **261**(5561), 552-558.

Li, Y., Han, L., Liu, Z., Wang, R. (2014a). Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *Journal of Chemical Information and Modeling,* **54**(6), 1717-1736.

Li, Y., Liu, Z., Li, J., Han, L., Liu, J., Zhao, Z., Wang, R. (2014b). Comparative assessment of scoring functions on an updated benchmark:

1. compilation of the test set. *Journal of Chemical Information and Modeling,* **54**(6), 1700-1716.

Liu, J., Montelione, G. T., Rost, B. (2007). Novel leverage of structural genomics. *Nature Biotechnology,* **25**(8), 849-851.

Liu, J., Wang, R. (2015). Classification of current scoring functions. *Journal of Chemical Information and Modeling,* **55**(3), 475-482.

Llobet, E., Campos, M. A., Gimenez, P., Moranta, D., Bengoechea, J. A. (2011). Analysis of the networks controlling the antimicrobial-peptide-dependent induction of klebsiella pneumoniae virulence factors. *Infection and Immunity,* **79**(9), 3718-3732.

Luthy, R., Bowie, J. U., Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature,* **356**(6364), 83-85.

Ma, B., Shatsky, M., Wolfson, H. J., Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Science,* **11**(2), 184-197.

Madej, T., Boguski, M. S., Bryant, S. H. (1995). Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Letters,* **373**(1), 13-18.

Maiti, R., Van Domselaar, G. H., Zhang, H., Wishart, D. S. (2004). SuperPose: A simple server for sophisticated structural superposition. *Nucleic Acids Research,* **32**(Web Server issue), W590-4.

Manabe, T., Kawano, M., Kawasaki, K. (2010). Mutations in the lipid A deacylase PagL which release the enzyme from its latency affect the ability of PagL to interact with lipopolysaccharide in salmonella enterica serovar typhimurium. *Biochemical and Biophysical Research Communications,* **396**(4), 812-816.

Manabe, T., Kawasaki, K. (2008). Extracellular loops of lipid A 3-O-deacylase PagL are involved in recognition of aminoarabinose-based membrane modifications in salmonella enterica serovar typhimurium. *Journal of Bacteriology,* **190**(16), 5597-5606.

Mancour, L. V., Daghestani, H. N., Dutta, S., Westfield, G. H., Schilling, J., Oleskie, A. N., Herbstman, J. F., Chou, S. Z., Skiniotis, G. (2012). Ligand-induced architecture of the leptin receptor signaling complex. *Molecular Cell,* **48**(4), 655-661.

Marceau, M. (2005). Transcriptional regulation in yersinia: An update. *Current Issues in Molecular Biology,* **7**(2), 151-177.

Marion, D. (2013). An introduction to biological NMR spectroscopy. *Molecular & Cellular Proteomics : MCP,* **12**(11), 3006-3025.

Marsden, R. L., Lewis, T. A., Orengo, C. A. (2007). Towards a comprehensive structural coverage of completed genomes: A structural genomics viewpoint. *BMC Bioinformatics,* **8**, 86.

Matsumura, I., Ellington, A. D. (2001). In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. *Journal of Molecular Biology,* **305**(2), 331-339.

McGinnis, K., Ku, G. M., VanDusen, W. J., Fu, J., Garsky, V., Stern, A. M., Friedman, P. A. (1996). Site-directed mutagenesis of residues in a conserved region of bovine aspartyl (asparaginyl) beta-hydroxylase: Evidence that histidine 675 has a role in binding Fe2+. *Biochemistry,* **35**(13), 3957-3962.

McGuffin, L. J. (2009). Prediction of global and local model quality in CASP8 using the ModFOLD server. *Proteins,* **77 Suppl 9**, 185-190.

McGuffin, L. J., Bryson, K., Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics,* **16**(4), 404-405.

McGuffin, L. J., Buenavista, M. T., Roche, D. B. (2013). The ModFOLD4 server for the quality assessment of 3D protein models. *Nucleic Acids Research,* **41**(Web Server issue), W368-72.

McGuffin, L. J., Roche, D. B. (2010). Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics,* **26**(2), 182-188.

McPherson, A. (2004). Introduction to protein crystallization. *Methods,* **34**(3), 254-265.

Merz, K. M., Ringe, D., & Reynolds, C. H. (2010). *Drug design: Structure- and ligand-based approaches*. New York, USA, Cambridge University Press.

Michino, M., Abola, E., GPCR Dock 2008 participants, Brooks, C. L.,3rd, Dixon, J. S., Moult, J., Stevens, R. C. (2009). Community-wide assessment of GPCR structure modelling and ligand docking: GPCR dock 2008. *Nature Reviews.Drug Discovery,* **8**(6), 455-463.

Mirabello, C., Pollastri, G. (2013). Porter, PaleAle 4.0: High-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics,* **29**(16), 2056-2058.

Moharana, K., Zabeau, L., Peelman, F., Ringler, P., Stahlberg, H., Tavernier, J., Savvides, S. N. (2014). Structural and mechanistic paradigm of leptin receptor activation revealed by complexes with wild-type and antagonist leptins. *Structure,* **22**(6), 866-877.

Montella, C., Bellsolell, L., Perez-Luque, R., Badia, J., Baldoma, L., Coll, M., Aguilar, J. (2005). Crystal structure of an iron-dependent group III dehydrogenase that interconverts L-lactaldehyde and L-1,2-propanediol in escherichia coli. *Journal of Bacteriology,* **187**(14), 4957-4966.

Montero, M., Eydallin, G., Viale, A. M., Almagro, G., Munoz, F. J., Rahimpour, M., Sesma, M. T., Baroja-Fernandez, E., Pozueta-Romero, J. (2009). Escherichia coli glycogen metabolism is controlled by the PhoP-PhoQ regulatory system at submillimolar environmental Mg2+ concentrations, and is highly interconnected with a wide variety of cellular processes. *The Biochemical Journal,* **424**(1), 129-141.

Monti, D., Ottolina, G., Carrea, G., Riva, S. (2011). Redox reactions catalyzed by isolated enzymes. *Chemical Reviews,* **111**(7), 4111-4140.

Morishita, E. C., Murayama, K., Kato-Murayama, M., Ishizuka-Katsura, Y., Tomabechi, Y., Hayashi, T., Terada, T., Handa, N., Shirouzu, M., Akiyama, T*., et al*. (2011). Crystal structures of the armadillo repeat domain of adenomatous polyposis coli and its complex with the tyrosine-rich domain of Sam68. *Structure,* **19**(10), 1496-1508.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins,* **82 Suppl 2**, 1-6.

Muegge, L., & Rarey, M. (2001). Small molecule docking and scoring. In K. B. Lipkowitz, & D. B. Boyd (Eds.), *Reviews in computational chemistry* (Vol 17 ed., pp. 1-60). New Jersey, Wiley-VCH Inc.

Mulder, N. J., Apweiler, R. (2002). Tools and resources for identifying protein families, domains and motifs. *Genome Biology,* **3**(1), REVIEWS2001.

Mullins, J. G. (2012). Structural modelling pipelines in next generation sequencing projects. *Advances in Protein Chemistry and Structural Biology,* **89**, 117-167.

Murata, T., Tseng, W., Guina, T., Miller, S. I., Nikaido, H. (2007). PhoPQ-mediated regulation produces a more robust permeability barrier in the outer membrane of salmonella enterica serovar typhimurium. *Journal of Bacteriology,* **189**(20), 7213-7222.

Murray, C. W., Auton, T. R., Eldridge, M. D. (1998). Empirical scoring functions. II. the testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of bayesian regression to improve the quality of the model. *Journal of Computer-Aided Molecular Design,* **12**(5), 503-519.

Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology,* **247**(4), 536-540.

Natt, N. K., Kaur, H., Raghava, G. P. (2004). Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins,* **56**(1), 11-18.

Navia, M. A., Fitzgerald, P. M., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L., Springer, J. P. (1989). Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature,* **337**(6208), 615-620.

Needleman, S. B., Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology,* **48**(3), 443-453.

Nelson, D. L., & Cox, M. M. (2005). *Lehninger principles of biochemistry* (4th ed.). New York, Freeman.

Neudert, G., Klebe, G. (2011). DSX: A knowledge-based scoring function for the assessment of protein-ligand complexes. *Journal of Chemical Information and Modeling,* **51**(10), 2731-2745.

Nichols, S. E., Baron, R., Ivetac, A., McCammon, J. A. (2011). Predictive power of molecular dynamics receptor structures in virtual screening. *Journal of Chemical Information and Modeling,* **51**(6), 1439-1446.

Niv-Spector, L., Gonen-Berger, D., Gourdou, I., Biener, E., Gussakovsky, E. E., Benomar, Y., Ramanujan, K. V., Taouis, M., Herman, B., Callebaut, I.*, et al*. (2005a). Identification of the hydrophobic strand in the A-B loop of leptin as major binding site III: Implications for large-scale preparation of potent recombinant human and ovine leptin antagonists. *The Biochemical Journal,* **391**(Pt 2), 221-230.

Niv-Spector, L., Raver, N., Friedman-Einat, M., Grosclaude, J., Gussakovsky, E. E., Livnah, O., Gertler, A. (2005b). Mapping leptin-interacting sites in recombinant leptin-binding domain (LBD) subcloned from chicken leptin receptor. *The Biochemical Journal,* **390**(Pt 2), 475-484.

Notredame, C., Higgins, D. G., Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology,* **302**(1), 205-217.

Novoa, E. M., de Pouplana, L. R., Barril, X., Orozco, M. (2010). The MM2QM tool for combining docking, molecular dynamics, molecular

mechanics, and quantum mechanics. *Journal of Chemical Theory and Computation,* **6**(8), 2547–2557.

Nowosielski, M., Hoffmann, M., Kuron, A., Korycka-Machala, M., Dziadek, J. (2013). The MM2QM tool for combining docking, molecular dynamics, molecular mechanics, and quantum mechanics. *Journal of Computational Chemistry,* **34**(9), 750-756.

Oertelt, C., Lindner, B., Skurnik, M., Holst, O. (2001). Isolation and structural characterization of an R-form lipopolysaccharide from yersinia enterocolitica serotype O:8. *European Journal of Biochemistry / FEBS,* **268**(3), 554-564.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure,* **5**(8), 1093-1108.

Orengo, C. A., Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. *Annual Review of Biochemistry,* **74**, 867-900.

Osguthorpe, D. J. (2000). Ab initio protein folding. *Current Opinion in Structural Biology,* **10**(2), 146-152.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., Notredame, C. (2004). 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology,* **340**(2), 385-395.

Overington, J. P., Al-Lazikani, B., Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews.Drug Discovery,* **5**(12), 993-996.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E.*, et al*. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science,* **289**(5480), 739-745.

Paliakasis, C. D., Michalopoulos, I., Kossida, S. (2008). Web-based tools for protein classification. *Methods in Molecular Biology (Clifton, N.J.),* **428**, 349-367.

Parmeggiani, F., Huang, P. S., Vorobiev, S., Xiao, R., Park, K., Caprari, S., Su, M., Seetharaman, J., Mao, L., Janjua, H.*, et al*. (2014). A general computational approach for repeat protein design. *Journal of Molecular Biology,* **427**(2), 563-575.

Parthier, C., Gorlich, S., Jaenecke, F., Breithaupt, C., Brauer, U., Fandrich, U., Clausnitzer, D., Wehmeier, U. F., Bottcher, C., Scheel, D.*, et al*. (2012). The O-carbamoyltransferase TobZ catalyzes an ancient enzymatic reaction. *Angewandte Chemie,* **51**(17), 4046-4052.

Pauling, L., Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America,* **37**(5), 251-256.

Pauling, L., Corey, R. B., Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America,* **37**(4), 205-211.

Pavlopoulou, A., Michalopoulos, I. (2011). State-of-the-art bioinformatics protein structure prediction tools (review). *International Journal of Molecular Medicine,* **28**(3), 295-310.

Pawlowski, M., Kozlowski, L., Kloczkowski, A. (2015). MQAPsingle: A quasi single-model approach for estimation of the quality of individual protein structure models. *Proteins,* doi: 10.1002/prot.24787.

Pearl, L. H. (1987). The catalytic mechanism of aspartic proteinases. *FEBS Letters,* **214**(1), 8-12.

Pearl, L. H., Blundell, T. (1984). The active site of aspartic proteinases. *FEBS Letters,* **174**(1), 96-101.

Pearl, L. H., Taylor, W. R. (1987). Sequence specificity of retroviral proteases. *Nature,* **328**(6130), 482.

Peelman, F., Zabeau, L., Moharana, K., Savvides, S. N., Tavernier, J. (2014). 20 years of leptin: Insights into signaling assemblies of the leptin receptor. *The Journal of Endocrinology,* **223**(1), T9-23.

Pei, J., Yin, N., Ma, X., Lai, L. (2014). Systems biology brings new dimensions for structure-based drug design. *Journal of the American Chemical Society,* **136**(33), 11556-11565.

Peitsch, M. C. (1996). ProMod and swiss-model: Internet-based tools for automated comparative protein modelling. *Biochemical Society Transactions,* **24**(1), 274-279.

Perez-Gutierrez, C., Llobet, E., Llompart, C. M., Reines, M., Bengoechea, J. A. (2010). Role of lipid A acylation in yersinia enterocolitica virulence. *Infection and Immunity,* **78**(6), 2768-2781.

Petkun, S., Shi, R., Li, Y., Asinas, A., Munger, C., Zhang, L., Waclawek, M., Soboh, B., Sawers, R. G., Cygler, M. (2011). Structure of hydrogenase maturation protein HypF with reaction intermediates shows two active sites. *Structure,* **19**(12), 1773-1783.

Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A.*, et al*. (2003). Using multiple structure alignments, fast model building, and

energetic analysis in fold recognition and homology modeling. *Proteins,* **53 Suppl 6**, 430-435.

Pillai, B., Cherney, M. M., Hiraga, K., Takada, K., Oda, K., James, M. N. (2007). Crystal structure of scytalidoglutamic peptidase with its first potent inhibitor provides insights into substrate specificity and catalysis. *Journal of Molecular Biology,* **365**(2), 343-361.

Pollastri, G., McLysaght, A. (2005). Porter: A new, accurate server for protein secondary structure prediction. *Bioinformatics,* **21**(8), 1719-1720.

Ponting, C. P., Russell, R. R. (2002). The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure,* **31**, 45-71.

Popovic, M., Sarngadharan, M. G., Read, E., Gallo, R. C. (1984). Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science,* **224**(4648), 497-500.

Power, M. D., Marx, P. A., Bryant, M. L., Gardner, M. B., Barr, P. J., Luciw, P. A. (1986). Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome retrovirus. *Science,* **231**(4745), 1567-1572.

Puustinen, P., Jaattela, M. (2014). KIAA1524/CIP2A promotes cancer growth by coordinating the activities of MTORC1 and MYC. *Autophagy,* **10**(7), 1352-1354.

Raetz, C. R. (1996). Bacterial lipopolysaccharides: A remarkable family of bioactive macroamphiphiles. In F. C. Neidhardt (Ed.), *Escherichia coli and salmonella: Cellular and molecular biology* (2nd, Vol 1. ed., pp. 1035-1063). Washington, DC, American Society for Microbiology.

Raetz, C. R. (1990). Biochemistry of endotoxins. *Annual Review of Biochemistry,* **59**, 129-170.

Raetz, C. R. (2001). Regulated covalent modifications of lipid A. *Journal of Endotoxin Research,* **7**(1), 73-78.

Raetz, C. R., Guan, Z., Ingram, B. O., Six, D. A., Song, F., Wang, X., Zhao, J. (2009). Discovery of new biosynthetic pathways: The lipid A story. *Journal of Lipid Research,* **50 Suppl**, S103-8.

Raetz, C. R., Reynolds, C. M., Trent, M. S., Bishop, R. E. (2007). Lipid A modification systems in gram-negative bacteria. *Annual Review of Biochemistry,* **76**, 295-329.

Raghava, G. P. S. (2002). APSSP2 : A combination method for protein secondary structure prediction based on neural network and example based learning. CASP5.**A-132.**

Ratner, L., Haseltine, W., Patarca, R., Livak, K. J., Starcich, B., Josephs, S. F., Doran, E. R., Rafalski, J. A., Whitehorn, E. A., Baumeister, K. (1985). Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature,* **313**(6000), 277-284.

Ray, A., Lindahl, E., Wallner, B. (2012). Improved model quality assessment using ProQ2. *BMC Bioinformatics,* **13**, 224-2105-13-224.

Read, R. J., Chavali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins,* **69 Suppl 8**, 27-37.

Rebeil, R., Ernst, R. K., Gowen, B. B., Miller, S. I., Hinnebusch, B. J. (2004). Variation in lipid A structure in the pathogenic yersiniae. *Molecular Microbiology,* **52**(5), 1363-1373.

Rebeil, R., Ernst, R. K., Jarrett, C. O., Adams, K. N., Miller, S. I., Hinnebusch, B. J. (2006). Characterization of late acyltransferase genes of yersinia pestis and their role in temperature-dependent lipid A variation. *Journal of Bacteriology,* **188**(4), 1381-1388.

Reichen, C., Hansen, S., Pluckthun, A. (2014). Modular peptide binding: From a comparison of natural binders to designed armadillo repeat proteins. *Journal of Structural Biology,* **185**(2), 147-162.

Reid, M. F., Fewson, C. A. (1994). Molecular characterization of microbial alcohol dehydrogenases. *Critical Reviews in Microbiology,* **20**(1), 13-56.

Reines, M., Llobet, E., Llompart, C. M., Moranta, D., Perez-Gutierrez, C., Bengoechea, J. A. (2012). Molecular basis of yersinia enterocolitica temperature-dependent resistance to antimicrobial peptides. *Journal of Bacteriology,* **194**(12), 3173-3188.

Remmert, M., Biegert, A., Hauser, A., Soding, J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods,* **9**(2), 173-175.

Reynolds, C. M., Ribeiro, A. A., McGrath, S. C., Cotter, R. J., Raetz, C. R., Trent, M. S. (2006). An outer membrane enzyme encoded by salmonella typhimurium lpxR that removes the 3'-acyloxyacyl moiety of lipid A. *The Journal of Biological Chemistry,* **281**(31), 21974-21987.

Rietschel, E. T., Kirikae, T., Schade, F. U., Mamat, U., Schmidt, G., Loppnow, H., Ulmer, A. J., Zahringer, U., Seydel, U., Di Padova, F.

(1994). Bacterial endotoxin: Molecular relationships of structure to activity and function. *FASEB Journal,* **8**(2), 217-225.

Rockah-Shmuel, L., Tawfik, D. S. (2012). Evolutionary transitions to new DNA methyltransferases through target site expansion and shrinkage. *Nucleic Acids Research,* **40**(22), 11627-11637.

Roman, N., Christie, M., Swarbrick, C. M., Kobe, B., Forwood, J. K. (2013). Structural characterisation of the nuclear import receptor importin alpha in complex with the bipartite NLS of Prp20. *PloS One,* **8**(12), e82038.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering,* **12**(2), 85-94.

Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology,* **318**(2), 595-608.

Roy, A., Kucukural, A., Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols,* **5**(4), 725-738.

Roy, A., Yang, J., Zhang, Y. (2012). COFACTOR: An accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research,* **40**(Web Server issue), W471-7.

Rutten, L., Mannie, J. P., Stead, C. M., Raetz, C. R., Reynolds, C. M., Bonvin, A. M., Tommassen, J. P., Egmond, M. R., Trent, M. S., Gros, P. (2009). Active-site architecture and catalytic mechanism of the lipid A deacylase LpxR of salmonella typhimurium. *Proceedings of the National Academy of Sciences of the United States of America,* **106**(6), 1960-1964.

Ryckaert, J., Ciccotti, G., Berendsen, H. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics,* **23**, 327.

Sadreyev, R. I., Grishin, N. V. (2004). Estimates of statistical significance for comparison of individual positions in multiple sequence alignments. *BMC Bioinformatics,* **5**, 106.

Sali, A. (1995). Comparative protein modeling by satisfaction of spatial restraints. *Molecular Medicine Today,* **1**(6), 270-277.

Sali, A., Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology,* **234**(3), 779-815.

Sali, A., Overington, J. P., Johnson, M. S., Blundell, T. L. (1990). From comparisons of protein sequences and structures to protein modelling and design. *Trends in Biochemical Sciences,* **15**(6), 235-240.

Salo, O. M., Raitio, K. H., Savinainen, J. R., Nevalainen, T., Lahtela-Kakkonen, M., Laitinen, J. T., Jarvinen, T., Poso, A. (2005). Virtual screening of novel CB2 ligands using a comparative model of the human cannabinoid CB2 receptor. *Journal of Medicinal Chemistry,* **48**(23), 7166-7171.

Sanchez, R., Sali, A. (1997). Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology,* **7**(2), 206-214.

Sanger, F. (1959). Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes. *Science,* **129**(3359), 1340-1344.

Schultz, J., Milpetz, F., Bork, P., Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America,* **95**(11), 5857-5864.

Schulz, G. E. (2000). Beta-barrel membrane proteins. *Current Opinion in Structural Biology,* **10**(4), 443-447.

Schulz, G. E. (2002). The structure of bacterial outer membrane proteins. *Biochimica Et Biophysica Acta,* **1565**(2), 308-317.

Schwede, T., Kopp, J., Guex, N., Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research,* **31**(13), 3381-3385.

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H. M., Jones, D., Brenner, S. E., Burley, S. K., Das, R., Dokholyan, N. V.*, et al*. (2009). Outcome of a workshop on applications of protein models in biomedical research. *Structure,* **17**(2), 151-159.

Setubal, J., & Meidanis, J. (1997). *Introduction to computational molecular biology* (1st ed.). Boston, PWS Publishing Company.

Shi, L., Javitch, J. A. (2002). The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop. *Annual Review of Pharmacology and Toxicology,* **42**, 437-467.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G.*, et al*. (2015). CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research,* **43**(Database issue), D376-81.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins,* **17**(4), 355-362.

Smith, R. D., Dunbar, J. B.,Jr, Ung, P. M., Esposito, E. X., Yang, C. Y., Wang, S., Carlson, H. A. (2011). CSAR benchmark exercise of 2010: Combined evaluation across all submitted scoring functions. *Journal of Chemical Information and Modeling,* **51**(9), 2115-2131.

Smith, T. F., Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology,* **147**(1), 195-197.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics,* **21**(7), 951-960.

Soding, J., Biegert, A., Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research,* **33**(Web Server issue), W244-8.

Sonnhammer, E. L., von Heijne, G., Krogh, A. (1998). A hidden markov model for predicting transmembrane helices in protein sequences. *Proceedings / International Conference on Intelligent Systems for Molecular Biology,* **6**, 175-182.

Soo Hoo, L., Zhang, J. Y., Chan, E. K. (2002). Cloning and characterization of a novel 90 kDa 'companion' auto-antigen of p62 overexpressed in cancer. *Oncogene,* **21**(32), 5006-5015.

Sousounis, K., Haney, C. E., Cao, J., Sunchu, B., Tsonis, P. A. (2012). Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Human Genomics,* **6**, 10-7364-6-10.

Sperandio, O., Mouawad, L., Pinto, E., Villoutreix, B. O., Perahia, D., Miteva, M. A. (2010). How to choose relevant multiple receptor conformations for virtual screening: A test case of Cdk2 and normal mode analysis. *European Biophysics Journal,* **39**(9), 1365-1372.

Stead, C. M., Beasley, A., Cotter, R. J., Trent, M. S. (2008). Deciphering the unusual acylation pattern of helicobacter pylori lipid A. *Journal of Bacteriology,* **190**(21), 7012-7021.

Straley, S. C., Perry, R. D. (1995). Environmental modulation of gene expression and pathogenesis in yersinia. *Trends in Microbiology,* **3**(8), 310-317.

Sun, J., Weis, W. I. (2011). Biochemical and structural characterization of beta-catenin interactions with nonphosphorylated and CK2-phosphorylated lef-1. *Journal of Molecular Biology,* **405**(2), 519-530.

Sutcliffe, M. J., Haneef, I., Carney, D., Blundell, T. L. (1987a). Knowledge based modelling of homologous proteins, part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engineering,* **1**(5), 377-384.

Sutcliffe, M. J., Hayes, F. R., Blundell, T. L. (1987b). Knowledge based modelling of homologous proteins, part II: Rules for the conformations of substituted sidechains. *Protein Engineering,* **1**(5), 385-392.

Tang, J., James, M. N., Hsu, I. N., Jenkins, J. A., Blundell, T. L. (1978). Structural evidence for gene duplication in the evolution of the acid proteases. *Nature,* **271**(5646), 618-621.

Tarendeau, F., Boudet, J., Guilligay, D., Mas, P. J., Bougault, C. M., Boulo, S., Baudin, F., Ruigrok, R. W., Daigle, N., Ellenberg, J*., et al*. (2007). Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nature Structural & Molecular Biology,* **14**(3), 229-233.

Tartaglia, L. A., Dembski, M., Weng, X., Deng, N., Culpepper, J., Devos, R., Richards, G. J., Campfield, L. A., Clark, F. T., Deeds, J*., et al*. (1995). Identification and expression cloning of a leptin receptor, OB-R. *Cell,* **83**(7), 1263-1271.

Teodoro, M. L., Kavraki, L. E. (2003). Conformational flexibility models for the receptor in structure based drug design. *Current Pharmaceutical Design,* **9**(20), 1635-1648.

Teodoro, M. L., Phillips, G. N.,Jr, Kavraki, L. E. (2003). Understanding protein flexibility through dimensionality reduction. *Journal of Computational Biology,* **10**(3-4), 617-634.

Teplova, M., Tereshko, V., Sanishvili, R., Joachimiak, A., Bushueva, T., Anderson, W. F., Egli, M. (2000). The structure of the yrdC gene product from escherichia coli reveals a new fold and suggests a role in RNA binding. *Protein Science,* **9**(12), 2557-2566.

Thorsteinsdottir, H. B., Schwede, T., Zoete, V., Meuwly, M. (2006). How inaccuracies in protein structure models affect estimates of protein-ligand interactions: Computational analysis of HIV-I protease inhibitor binding. *Proteins,* **65**(2), 407-423.

Tian, W., Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology,* **333**(4), 863-882.

Tikhonova, I. G., Sum, C. S., Neumann, S., Engel, S., Raaka, B. M., Costanzi, S., Gershengorn, M. C. (2008). Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *Journal of Medicinal Chemistry,* **51**(3), 625-633.

Todd, A. E., Marsden, R. L., Thornton, J. M., Orengo, C. A. (2005). Progress of structural genomics initiatives: An analysis of solved target structures. *Journal of Molecular Biology,* **348**(5), 1235-1260.

Todd, A. E., Orengo, C. A., Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology,* **307**(4), 1113-1143.

Toh, H., Hayashida, H., Kikuno, R., Yasunaga, T., Miyata, T. (1985). Sequence similarity between EGF receptor and alpha 1-acid glycoprotein. *Nature,* **314**(6007), 199.

Toivola, J., Nikkanen, L., Dahlstrom, K. M., Salminen, T. A., Lepisto, A., Vignols, H. F., Rintamaki, E. (2013). Overexpression of chloroplast NADPH-dependent thioredoxin reductase in arabidopsis enhances leaf growth and elucidates in vivo function of reductase and thioredoxin domains. *Frontiers in Plant Science,* **4**, 389.

Tou, W. I., Chang, S. S., Lee, C. C., Chen, C. Y. (2013). Drug design for neuropathic pain regulation from traditional chinese medicine. *Scientific Reports,* **3**, 844.

Tou, W. I., Chen, C. Y. (2014). May disordered protein cause serious drug side effect? *Drug Discovery Today,* **19**(4), 367-372.

Tracewell, C. A., Arnold, F. H. (2009). Directed enzyme evolution: Climbing fitness peaks one amino acid at a time. *Current Opinion in Chemical Biology,* **13**(1), 3-9.

Tramontano, A., Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins,* **53 Suppl 6**, 352-368.

Tramontano, A. (2006). *Protein structure prediction: Concepts and applications*. Weinheim, Wiley-VCH.

Trent, M. S., Stead, C. M., Tran, A. X., Hankins, J. V. (2006). Diversity of endotoxin and its impact on pathogenesis. *Journal of Endotoxin Research,* **12**(4), 205-223.

Tress, M. L., Grana, O., Valencia, A. (2004). SQUARE--determining reliable regions in sequence alignments. *Bioinformatics,* **20**(6), 974-975.

Tress, M. L., Jones, D., Valencia, A. (2003). Predicting reliable regions in protein alignments from sequence profiles. *Journal of Molecular Biology,* **330**(4), 705-718.

Tsai, C. J., Kumar, S., Ma, B., Nussinov, R. (1999a). Folding funnels, binding funnels, and protein function. *Protein Science,* **8**(6), 1181-1190.

Tsai, C. J., Ma, B., Nussinov, R. (1999b). Folding and binding cascades: Shifts in energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America,* **96**(18), 9970-9972.

UniProt Consortium. (2015). UniProt: A hub for protein information. *Nucleic Acids Research,* **43**(Database issue), D204-12.

Vangrevelinghe, E., Zimmermann, K., Schoepfer, J., Portmann, R., Fabbro, D., Furet, P. (2003). Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *Journal of Medicinal Chemistry,* **46**(13), 2656-2662.

Varadamsetty, G., Tremmel, D., Hansen, S., Parmeggiani, F., Pluckthun, A. (2012). Designed armadillo repeat proteins: Library generation, characterization and selection of peptide binders with high specificity. *Journal of Molecular Biology,* **424**(1-2), 68-87.

Velec, H. F., Gohlke, H., Klebe, G. (2005). DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *Journal of Medicinal Chemistry,* **48**(20), 6296-6303.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A.*, et al*. (2001). The sequence of the human genome. *Science,* **291**(5507), 1304-1351.

Verkhivker, G., Appelt, K., Freer, S. T., Villafranca, J. E. (1995). Empirical free energy calculations of ligand-protein crystallographic complexes. I. knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Engineering,* **8**(7), 677-691.

von Heijne, G. (1992). Membrane protein structure prediction. hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology,* **225**(2), 487-494.

von Heijne, G. (1996). Principles of membrane protein assembly and structure. *Progress in Biophysics and Molecular Biology,* **66**(2), 113-139.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., Notredame, C. (2006). M-coffee: Combining multiple sequence alignment methods with T-coffee. *Nucleic Acids Research,* **34**(6), 1692-1699.

Wallner, B., Elofsson, A. (2003). Can correct protein models be identified? *Protein Science,* **12**(5), 1073-1086.

Wallner, B., Elofsson, A. (2005). All are not equal: A benchmark of different homology modeling programs. *Protein Science,* **14**(5), 1315-1327.

Wallner, B., Elofsson, A. (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Science,* **15**(4), 900-913.

Wang, Q., Vantasin, K., Xu, D., Shang, Y. (2011). MUFOLD-WQA: A new selective consensus method for quality assessment in protein structure prediction. *Proteins,* **79 Suppl 10**, 185-195.

Wang, Z., Cheng, J. (2012). An iterative self-refining and self-evaluating approach for protein model quality estimation. *Protein Science,* **21**(1), 142-151.

Wang, Z., Eickholt, J., Cheng, J. (2011). APOLLO: A quality assessment service for single and multiple protein models. *Bioinformatics,* **27**(12), 1715-1716.

Warren, G. L., Do, T. D., Kelley, B. P., Nicholls, A., Warren, S. D. (2012). Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today,* **17**(23-24), 1270-1281.

Wheeler, D. (2002). Selecting the right protein-scoring matrix. *Current Protocols in Bioinformatics / Editoral Board, Andreas D.Baxevanis et al.,* **Chapter 3**, Unit 3.5.

Wheeler, D., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A*., et al*. (2003). Database resources of the national center for biotechnology. *Nucleic Acids Research,* **31**(1), 28-33.

Wiederstein, M., Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research,* **35**(Web Server issue), W407-10.

Wilmot, C. M., Hajdu, J., McPherson, M. J., Knowles, P. F., Phillips, S. E. (1999). Visualization of dioxygen bound to copper during enzyme catalysis. *Science,* **286**(5445), 1724-1728.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J., Kent, S. B. (1989). Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. *Science,* **245**(4918), 616-621.

Wu, C. H., Huang, H., Yeh, L. S., Barker, W. C. (2003). Protein family classification and functional annotation. *Computational Biology and Chemistry,* **27**(1), 37-47.

Wu, S., Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research,* **35**(10), 3375-3382.

Wu, Y., Gu, T. T., Zheng, P. S. (2015). CIP2A cooperates with H-ras to promote epithelial-mesenchymal transition in cervical-cancer progression. *Cancer Letters,* **356**(2 Pt B), 646-655.

Xiang, Z., Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. *Journal of Molecular Biology,* **311**(2), 421-430.

Xiang, Z., Soto, C. S., Honig, B. (2002). Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences of the United States of America,* **99**(11), 7432-7437.

Xu, W., Kimelman, D. (2007). Mechanistic insights from structural studies of beta-catenin and its binding partners. *Journal of Cell Science,* **120**(Pt 19), 3337-3344.

Xue, Y., Wu, G., Wang, X., Zou, X., Zhang, G., Xiao, R., Yuan, Y., Long, D., Yang, J., Wu, Y*., et al*. (2013). CIP2A is a predictor of survival and a novel therapeutic target in bladder urothelial cell carcinoma. *Medical Oncology,* **30**(1), 406-012-0406-6. Epub 2012 Dec 30.

Yang, J., Roy, A., Zhang, Y. (2013). BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research,* **41**(Database issue), D1096-103.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y. (2015). The I-TASSER suite: Protein structure and function prediction. *Nature Methods,* **12**(1), 7-8.

Yang, Y., Zhou, Y. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Science,* **17**(7), 1212-1219.

Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G. M., Bhattacharyya, S., Gutierrez, P*., et al*. (2002). An NMR approach to structural proteomics. *Proceedings of the National Academy of Sciences of the United States of America,* **99**(4), 1825-1830.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research,* **31**(13), 3370-3374.

Zemla, A., Venclovas, C., Moult, J., Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins,* **Suppl 3**, 22-29.

Zemla, A., Venclovas, Moult, J., Fidelis, K. (2001). Processing and evaluation of predictions in CASP4. *Proteins,* **Suppl 5**, 13-21.

Zhang, F., Basinski, M. B., Beals, J. M., Briggs, S. L., Churgay, L. M., Clawson, D. K., DiMarchi, R. D., Furman, T. C., Hale, J. E., Hsiung, H.

M., *et al*. (1997). Crystal structure of the obese protein leptin-E100. *Nature,* **387**(6629), 206-209.

Zhang, Y. (2008a). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology,* **18**(3), 342-348.

Zhang, Y. (2008b). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics,* **9**, 40-2105-9-40.

Zhang, Y. (2014). Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins,* **82 Suppl 2**, 175-187.

Zhang, Y., Arakaki, A. K., Skolnick, J. (2005). TASSER: An automated method for the prediction of protein tertiary structures in CASP6. *Proteins,* **61 Suppl 7**, 91-98.

Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L., Friedman, J. M. (1994). Positional cloning of the mouse obese gene and its human homologue. *Nature,* **372**(6505), 425-432.

Zhang, Z., Li, Y., Lin, B., Schroeder, M., Huang, B. (2011). Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics,* **27**(15), 2083-2088.

Zheng, M., Liu, X., Xu, Y., Li, H., Luo, C., Jiang, H. (2013). Computational methods for drug design and discovery: Focus on china. *Trends in Pharmacological Sciences,* **34**(10), 549-559.

ISBN 978-952-12-3299-2