

Linköping studies in science and technology  
Dissertation No. 1271

# Mutational effects on protein structure and function

Jonas Carlsson



**Linköping University**  
**INSTITUTE OF TECHNOLOGY**

Department of Physics, Chemistry and Biology  
Linköping university  
SE-581 83 Linköping, Sweden

Linköping 2009

The cover picture shows the model structure of steroid 21-hydroxylase with bound heme co-factor in red and bound steroid ligand progesterone in green. Mutations with varying clinical disease phenotype, found in human patients, are indicated in the structure with the following color coding: red corresponds to SW, blue to SV and yellow to NC.

During the course of the research underlying this thesis, Jonas Carlsson was enrolled in Forum Scientium, a multidisciplinary doctoral programme at Linköping University, Sweden.

Copyright © Jonas Carlsson 2009, unless otherwise noted.  
All rights reserved.

**Mutational effects on protein structure and function**

ISBN 978-91-7393-539-5

ISSN 0345-7524

Printed by LiU-Tryck, Linköping, Sweden 2009

# Abstract

In this thesis several important proteins are investigated from a structural perspective. Some of the proteins are disease related while other have important but not completely characterised functions. The techniques used are general as demonstrated by applications on metabolic proteins (CYP21, CYP11B1, IAPP, ADH3), regulatory proteins (p53, GDNF) and a transporter protein (ANTR1).

When the protein CYP21 (steroid 21-hydroxylase) is deficient it causes CAH (congenital adrenal hyperplasia). For this protein, there are about 60 known mutations with characterised clinical phenotypes. Using manual structural analysis we managed to explain the severity of all but one of the mutations. By observing the properties of these mutations we could perform good predictions on, at the time, not classified mutations.

For the cancer suppressor protein p53, there are over thousand mutations with known activity. To be able to analyse such a large number of mutations we developed an automated method for evaluation of the mutation effect called PRED-MUT. In this method we include twelve different prediction parameters including two energy parameters calculated using an energy minimization procedure. The method manages to differentiate severe mutations from non-severe mutations with 77% accuracy on all possible single base substitutions and with 88% on mutations found in breast cancer patients.

The automated prediction was further applied to CYP11B1 (steroid 11-beta-hydroxylase), which in a similar way as CYP21 causes CAH when deficient. A generalized method applicable to any kind of globular protein was developed. The method was subsequently evaluated on nine additional proteins for which mutants were known with annotated disease phenotypes. This prediction achieved 84% accuracy on CYP11B1 and 81% accuracy in total on the evaluation proteins while leaving 8% as unclassified. By increasing the number of unclassified mutations the accuracy of the remaining mutations could be increased on the evaluation proteins and substantially increase the classification quality as measured by the Matthews correlation coefficient. Servers with predictions for all possible single based substitutions are provided for p53, CYP21 and CYP11B1.

The amyloid formation of IAPP (islet amyloid polypeptide) is strongly connected to diabetes and has been studied using both molecular dynamics and Monte Carlo energy minimization. The effects of mutations on the amount and speed of amyloid formation were investigated using three approaches. Applying a consensus of the three methods on a number of interesting mutations, 94% of the mutations could be correctly classified as amyloid forming or not, evaluated with *in vitro*

measurements.

In the brain there are many proteins whose functions and interactions are largely unknown. GDNF (glial cell line-derived neurotrophic factor) and NCAM (neural cell adhesion molecule) are two such neuron connected proteins that are known to interact. The form of interaction was studied using protein–protein docking where a docking interface was found mediated by four oppositely charged residues in respective protein. This interface was subsequently confirmed by mutagenesis experiments. The NCAM dimer interface upon binding to the GDNF dimer was also mapped as well as an additional interacting protein, GFR $\alpha$ 1, which was successfully added to the protein complex without any clashes.

A large and well studied protein family is the alcohol dehydrogenase family, ADH. A class of this family is ADH3 (alcohol dehydrogenase class III) that has several known substrates and inhibitors. By using virtual screening we tried to characterize new ligands. As some ligands were already known we could incorporate this knowledge when the compound docking simulations were scored and thereby find two new substrates and two new inhibitors which were subsequently successfully tested *in vitro*.

ANTR1 (anion transporter 1) is a membrane bound transporter important in the photosynthesis in plants. To be able to study the amino acid residues involved in inorganic phosphate transportation a homology model of the protein was created. Important residues were then mapped onto the structure using conservation analysis and we were in this way able to propose roles of amino acid residues involved in the transportation of inorganic phosphate. Key residues were subsequently mutated *in vitro* and a transportation process could be postulated.

To conclude, we have used several molecular modelling techniques to find functional clues, interaction sites and new ligands. Furthermore, we have investigated the effect of mutations on the function and structure of a multitude of disease related proteins.

# Populärvetenskaplig sammanfattning

Denna avhandling är främst en studie av hur små förändringar i ett protein kan leda till olika typer av sjukdomar. För att lyckas med detta har många vetenskapliga discipliner sammanflätats. Inom biokemin studerar man proteiners funktion och struktur på labb och i djur. I medicinsk forskning diagnostiserar man patienters kliniska tillstånd och kategoriserar vilka proteiner det är fel på. Inom bioinformatik samlas biologisk information först ihop och sedan analyseras. Till slut har vi beräkningsfysik, vars forskning har möjliggjort simuleringar av t ex proteiner med hjälp av datorer. Kunskap och kompetens inom alla dessa områden har behövts i de projekt jag har varit inblandad i.

Proteiner är egentligen ganska enkla. De är uppbyggda av 20 olika byggstenar, s.k. aminosyror, som när de är hopsatta kan liknas vid ett långt snöre. På grund av att byggstenarna har olika egenskaper, som att de har laddningar eller är feta, kommer delar av proteinet att dras till varandra. Minusladdade aminosyror dras till plusladdade och feta aminosyrorna dras till andra feta ungefär som olja i vatten eftersom en cell mestadels består av vatten. Denna attraktionskraft mellan aminosyrorna gör att proteinet kommer att rulla ihop sig likt ett nystan, men inte vilket nystan som helst, utan ett välordnat nystan som ser likadant ut varje gång. Denna unika struktur gör det möjligt för olika proteiner att utföra olika uppgifter.

Ett protein har ofta en relativt enkel uppgift, t ex att modifiera en del av en viss molekyl genom att exempelvis ta bort en väteatom från en kolatom och istället sätta dit en syreatom. Detta förändrar molekylens egenskaper vilket kan vara av nytta för olika processer som ständigt sker i vår kropp. Den del av proteinet där denna modifiering sker brukar kallas för den aktiva ytan och består oftast av en liten grotta i proteinet där molekylerna passar in. Några aminosyror i proteinet håller fast molekylerna medan några andra utför modifieringen. Förändringar av molekyler sker hela tiden spontant men ofta ganska långsamt och den egentliga uppgiften för ett protein är att snabba upp denna process. Man säger att proteinet är en katalysator för reaktionen och genom sin speciella struktur ser proteinet till att alla förutsättningar är optimala för att just denna reaktion ska ske.

I kroppen finns över 20000 olika proteiner som hela tiden utför var sin uppgift och dessa tillsammans gör att kroppen fungerar. Så vad skulle hända om man förändrar en liten byggsten i ett av dessa 20000 proteiner. Ja, det är förstås beroende på vilket protein som drabbas, vilken sorts förändring det är samt var i proteinet förändringen sker. Vissa proteiner har dessutom backupsystem som gör att andra proteiner helt eller delvis kan ta över det icke fungerande proteinets roll. Det har också stor betydelse om förändringen är medfödd eller om den uppstår i en enskild cell. En allvarlig medfödd störning av ett protein kan leda till döden medan ett icke fungerande protein i en cell kanske bara medför att cellen dör eller fungerar lite sämre än de andra. Problem kan uppstå då förändringar sker i de proteiner som ansvarar för celledningen. Då ökar risken för cancer.

Proteinets funktion är beroende av den väldefinierade strukturen, både generellt och i den aktiva ytan. Alla förändringar gör dock inte att proteinet slutar fungera. Byter man t ex ut en plusladdad aminosyra mot en annan plusladdad är det ganska stor chans att strukturen är bevarad. Byter man istället ut den plusladdade mot en oladdad aminosyra kan strukturens stabilitet minska vilket gör att proteinets funktion försämras eller i värsta fall helt förstörs. Den vanligaste förändringen som helt tar bort proteinets funktion är förändringar i aktiva ytan. Byter man ut de funktionella aminosyrorna så försämras förutsättningarna för att reaktionen ska äga rum, ofta så pass mycket att proteinet slutar fungera.

Genom att studera de olika egenskaperna hos aminosyrorna i kombination med annan kunskap om proteinet kan man försöka förutsäga vilka förändringar i proteinet som leder till vilka följder, både för proteinet och för en patient. Detta är vad jag har gjort för ett flertal protein som finns i människan. Jag har bl a studerat p53 som är vårt viktigaste protein i kampen mot cancer. Det upptäcker nämligen DNA-skador och ser till att dessa fel rättas till, alternativt ser till att cellen dör för att skydda resten av kroppen.

En patient som har ett muterat protein drabbas olika mycket beroende på hur stor effekt förändringen har på proteins funktion. Sjukdomsgraden klassas ofta in i olika fenotyper, som kräver olika mycket behandling. Jag har utvecklat ett generellt system som man kan använda som hjälpmedel för att förutsäga och diagnostisera sjukdomsfenotyper hos patienter där man känner till vilken sorts förändring en patient har av det aktuella proteinet. Kravet är dock att dess tredimensionella struktur är känd. Genom att förutse effekten av en patients mutation är det möjligt att hjälpa till med sjukdomsklassificering och då finns möjligheten att fler patienter tidigt kan få en korrekt behandling. Kunskapen om vad som gör att ett protein inte fungerar kan även användas för att komma ett steg närmare till att hitta nya och effektivare läkemedel.

Hittills har vi bara berört förändringar i proteinets struktur och funktion genom mutationer. Ibland behövs det inga mutationer för att proteinet ska förlora sin funktion och en sjukdom ska uppstå. Protein kan nämligen spontant vecka sig felaktigt och klumpa ihop sig med varandra. Detta uppträder relativt ofta i vissa

---

proteiner medan det i de flesta är ganska ovanligt. Ett protein som ofta drabbas av felveckningar är IAPP, vilket är ett protein som reglerar insulinproduktionen och produceras av samma celler som producerar insulin. Om en IAPP-molekyl veckar sig felaktigt och kommer i kontakt med en annan IAPP-molekyl kan denna också felveckas eftersom detta skapar fördelaktiga interaktioner mellan proteinmolekylerna. Detta komplex är väldigt stabilt och kommer i ytterkanterna att omvandla fler molekyler vilket medför att en växande fiber av IAPP-molekyler bildas. När dessa komplex är små förgiftar de cellerna de produceras i och när de har växt sig riktigt stora kan de störa cellens funktion genom att helt enkelt vara i vägen. När de celler som producerar insulin och IAPP dör minskar insulinproduktionen vilket är en vanlig orsak till att man får diabetes av typ II, s.k. åldersdiabetes.

Genom datorsimuleringar av IAPP har vi kunnat studera aspekter av hur felveckningen går till, vilka delar som är viktiga för felveckningen samt förutsäga vilka förändringar i proteinstrukturen som motverkar felveckning och den resulterande fiberbildningen.

Det kan även vara intressant att studera proteiner som fungerar som de ska. Uppgiften är då istället att ta reda på så mycket som möjligt om proteinet som inte redan är känt, t ex genom att hitta nya molekyler som proteinet är verksamt på, s.k. substrat. Mha datorsimuleringar kan man testa stora bibliotek av molekyler och se hur bra de passar i proteinets aktiva yta, vilket vi gjorde på ett protein som heter alkoholdehydrogenas klass III och fann både nya substrat och inhibitorer. Man kan också undersöka samspelet med andra proteiner vilket vi gjorde i ett projekt involverande två proteiner som är aktiva i hjärnan. I ytterligare ett projekt har vi studerat transportmekanismen hos ett membranbundet protein.

Sammanfattningsvis har vi använt datorsimuleringar som ett redskap för att beskriva proteiners naturliga processer samt förändringar som kan leda till minskad enzymaktivitet och i förlängningen ge sjukdomar.



# Acknowledgements

During my thesis there have been many people that have helped me, either in a scientific way, in support, or by making my time at the university a very pleasant and interesting time.

The biggest contributor to the science part of my thesis is of course my supervisor Bengt. His inputs have been invaluable with his surprising number of ideas and positive thinking. Even though he lives in Stockholm he has always been easy to get in contact with, even on vacations, and has in total travelled 5 times around the earth just to lead our little group of three PhD students.

The bioinformatics group has always been tight. Joel has been my room mate for almost four years. We have always had fun together; we have listened to good and bad hard rock music, told funny and not so funny jokes, had scientific discussions and made geeky computer stuff.

Dr. Anders left the building and the city quite a while ago, but before that was part of the group. He is a good friend that always was there for you when you needed it. He is also so enthusiastic about things that even a dull football match could be interesting to watch in his presence.

The last person in our group is Fredrik who is a relative newcomer. He is a person that enjoys life and challenges in such a way that the positive energy surrounding him can make your hardest problems look like child's play.

I would also like to thank all my collaborators, especially Tiina Robins, Anna Wedell, Mikko Hellgren, Dan Sjöstrand, Gunilla Westermarck and Thierry Soussi. Without your help my thesis would be much thinner and my time in Linköping much less interesting.

Science is one thing, but to produce good science you have to feel good. And what can make you feel good if not for your friends. I would especially thank Chrul for being the nicest guy in the world, Kristofer for giving me the first really good knick name, Erik for making me get lost in the woods, and Janosch for letting me win in tennis.

Other important people at IFM or previously at IFM are Pelle, Peter, Fredrik,

Andréas, Jenny, Andreas, Neda, Patrick, Tom, Olle and several more that has made my time here more enjoyable. My appreciation goes also to Stefan for heading the superb arrangement of Forum Scientium.

There are some old friends of mine that I would like to mention. Particularly, Gary that I lived together with for a long time and Lisa that did stand to live with us both. Friends from my study time that deserves special mention are Jinnis, Peter, Magni, Brynis, Luc, Hedis and Miro. I would also like show my appreciation to Urban and Karin for letting us stay at "Hovet" and for being such nice friends and Andreas and Jennie just for being who you are.

Last but not least I would like to thank my family. Special gratitude to my parents who have helped me with the thesis, mental support, but mostly for by being my best parents I have ever had! Final words goes to my Karin for being the best person in the entire universe, at everything and in every real and imaginary way possible, for believing in me, for loving me, for ever.

# List of publications

## Paper I

Molecular model of human CYP21 based on mammalian CYP2C5: structural features correlate with clinical severity of mutations causing congenital adrenal hyperplasia

Robins, T.\*, Carlsson, J.\*, Sunnerhagen, M., Wedell, A., Persson, B.

\*Shared first authors

*Mol Endocrinol*, 20:2946–2964, 2006, PMID: 16788163

## Paper II

Investigation and prediction of the severity of p53 mutants using parameters from structural calculations

Jonas Carlsson, Thierry Soussi, Bengt Persson

*FEBS Journal*, 276:4142–4155, 2009, PMID: 19558493

## Paper III

A structural model of human steroid 11-beta-hydroxylase, CYP11B1, used to predict consequences of mutations

Jonas Carlsson, Anna Wedell, Bengt Persson

*Adv in Bioinformatics*, 2009, Submitted

## Paper IV

Disruption of the GDNF binding site in NCAM dissociates ligand binding and homophilic cell adhesion

Dan Sjöstrand, Jonas Carlsson, Gustavo Paratcha, Bengt Persson, Carlos F. Ibáñez

*J Biol Chem*, 282:12734–12740, 2007, PMID: 17322291

## Paper V

Functionally Important Amino Acids in the Arabidopsis Thylakoid Phosphate Transporter – Homology Modeling and Site-directed Mutagenesis

Lorena Ruiz Pavón, Patrik Karlsson, Jonas Carlsson, Dieter Samyn, Bengt Persson, Bengt L. Persson, Cornelia Spetea

*In manuscript*

**Paper VI**

A folding study on IAPP (Islet Amyloid Polypeptide) using molecular dynamics simulations

Jonas Carlsson, Aida Vahdat Shariatpanahi, Sebastian Schultz, Gunilla T. Westermarck, Bengt Persson

*In manuscript*

**Paper VII**

Virtual screening for ligands to human alcohol dehydrogenase 3

Mikko Hellgren, Jonas Carlsson, Linus Östberg, Claudia A. Staab, Bengt Persson, Jan-Olov Höög

*In manuscript*

**Publications not included in the thesis****Paper SI**

Unfolding a folding disease: folding, misfolding and aggregation of the marble brain syndrome-associated mutant H107Y of human carbonic anhydrase II

Karin Almstedt, Martin Lundqvist, Jonas Carlsson, Martin Karlsson, Bengt Persson, Bengt-Harald Jonsson, Uno Carlsson, Per Hammarstrom

*J Mol Biol*, 342:619–633, 2004, PMID: 15327960

**Paper SII**

A promiscuous glutathione transferase transformed into a selective thiolester hydrolase

Sofia Hederos, Lotta Tegler, Jonas Carlsson, Bengt Persson, Johan Viljanen, Kerstin S. Broo

*Org Biomol Chem*, 4:90–97, 2006, PMID: 16358001

**Paper SIII**

Three novel CYP21A2 mutations and their protein modelling in patients with classical 21-hydroxylase deficiency from northeastern Iran

Alireza Baradaran-Heravi, Rahim Vakili, Tiina Robins, Jonas Carlsson, Nosrat Ghaemi, Azadeh A'Rabi, Mohammad Reza Abbaszadegan

*Mol Endocrinol*, 67:335–341, 2007, PMID: 17573904

**Paper SIV**

A new polymorphism in the coding region of exon four in HSD17B2 in relation to risk of sporadic and hereditary breast cancer

Agneta Jansson, Jonas Carlsson, Anette Olsson, Petter Storm, Sara Margolin, Cecilia Gunnarsson, Marie Stenmark-Askmal, Annika Lindblom, Bengt Persson, Olle Stål

*Breast Cancer Res Treat*, 106:57–64, 2007, PMID: 17260097

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Bioinformatics . . . . .	1
1.2	Proteins . . . . .	1
1.2.1	Protein regulation . . . . .	2
1.2.2	Genes . . . . .	3
1.3	Mutations . . . . .	3
<b>2</b>	<b>Methods</b>	<b>5</b>
2.1	Molecular modelling . . . . .	5
2.1.1	Visualization . . . . .	5
2.1.2	Protein folding . . . . .	6
2.1.3	Energy minimization . . . . .	6
2.1.4	Monte Carlo based energy minimization . . . . .	7
2.1.5	Molecular dynamics . . . . .	9
2.1.6	Energy terms . . . . .	11
2.1.7	Force fields . . . . .	14
2.1.8	CASP . . . . .	14
2.2	Docking . . . . .	15
2.2.1	Protein–protein docking . . . . .	15
2.2.2	Protein–ligand docking . . . . .	16
2.2.3	Virtual ligand screening . . . . .	16
2.3	Superimposition . . . . .	18
2.4	Homology modelling . . . . .	20
2.4.1	Preparation . . . . .	21
2.4.2	Evaluation . . . . .	22
2.5	<i>Ab initio</i> modelling . . . . .	24
2.6	Secondary structure prediction . . . . .	24
2.7	Methods for prediction . . . . .	26
2.7.1	Monte Carlo . . . . .	26
2.7.2	Principal component analysis . . . . .	26
2.7.3	Support vector machines . . . . .	27

---

2.7.4	Decision trees . . . . .	28
2.7.5	Consensus . . . . .	29
2.7.6	Evaluation . . . . .	29
2.8	Protein structure databases . . . . .	31
2.9	Tools . . . . .	32
2.9.1	Molecular modelling . . . . .	32
2.9.2	Homology modelling . . . . .	32
2.9.3	<i>Ab initio</i> modelling . . . . .	33
2.9.4	Docking . . . . .	33
2.9.5	Statistical analysis . . . . .	34
2.9.6	Mutation evaluation servers . . . . .	34
<b>3</b>	<b>Studied proteins</b>	<b>37</b>
3.1	Introduction . . . . .	37
3.2	Motivation . . . . .	38
3.3	Steroid 21-hydroxylase . . . . .	38
3.4	Steroid 11 $\beta$ -hydroxylase . . . . .	39
3.5	p53 . . . . .	39
3.6	Islet Amyloid Polypeptide . . . . .	40
3.7	Neural Cell Adhesion Molecule . . . . .	41
3.8	Glial cell derived neurotrophic factor . . . . .	41
3.9	The anion transporter 1 . . . . .	42
3.10	Alcohol dehydrogenase class III . . . . .	43
<b>4</b>	<b>Summary of papers</b>	<b>45</b>
4.1	Paper I . . . . .	45
4.2	Paper II . . . . .	47
4.3	Paper III . . . . .	49
4.4	Paper IV . . . . .	50
4.5	Paper V . . . . .	52
4.6	Paper VI . . . . .	53
4.7	Paper VII . . . . .	55
<b>5</b>	<b>Discussion of results</b>	<b>57</b>
5.1	New mutations in CYP21 . . . . .	57
5.2	Mutations . . . . .	57
5.2.1	Effect on stability . . . . .	58
5.2.2	Effect on function . . . . .	58
5.2.3	Measurable variables . . . . .	59
5.3	Future ideas . . . . .	61
5.3.1	Additional prediction parameters . . . . .	61
5.3.2	Evaluation of some additional parameters . . . . .	62
5.3.3	A new prediction strategy . . . . .	63

5.3.4	Consensus . . . . .	63
5.3.5	Future outlook . . . . .	64
<b>6</b>	<b>Conclusions</b>	<b>67</b>
	<b>Bibliography</b>	<b>69</b>
	<b>Publications</b>	<b>81</b>



# Chapter 1

## Introduction

My research area is bioinformatics where I have studied the effects of amino acid changes upon function and structure in different proteins found mostly in humans. Several of these proteins are strongly connected to common diseases that affect millions of people worldwide.

### 1.1 Bioinformatics

The field of bioinformatics is very broad. Computers are used to investigate information assembled by biologists all over the world. One part of bioinformatics is to manage databases where biological information is stored and provide infrastructure for searches and management of these databases. Another part is the analysis of the existing biological data. The information concerns different levels in the cell, *e.g.* DNA, genes, proteins, interactions and reactions. My area of interest lies mostly on the protein level, even if it is necessary and important to involve information from different levels.

### 1.2 Proteins

Proteins are like small machines or tools in the body that perform most of the tasks needed to be done in order for us to function. The proteins often work together, sometimes as a complex and other times separately but in a consecutive order. The proteins themselves consist of 20 different types of building blocks, amino acids. The amino acids are connected via a peptide bond that forms the main chain of the protein, rather like yarn. What make the amino acids different from each other are the side chains that have different sizes and chemical properties. Depending on the properties of the side chains, each protein main chain will fold into a specific configuration every time. For cytosolic proteins, the folding

process is mainly governed by the fact that the hydrophobic amino acid residues group together as they minimize the accessible surface to the surrounding water in addition to formation of favourable hydrogen bonds both in the main chain, called secondary structures, and between side chains. In the case of membrane proteins the surrounding environment is inverted as the membrane itself is made out of lipids and the hydrophobic residues will instead tend to be located on the surface while the hydrophilic residues will mostly be in the core or on in the loops outside of the membrane.

Once the protein has folded it can start to perform its function. Many proteins can perform different tasks, however, the protein often has a favourite molecule that it can change properties of or transport most effectively. Enzymes, that modify a chemical bond or substructure in a molecule, have a binding pocket where the substrate fits in a specific way. The part of the bound molecule that is going to be changed is located on an optimal distance to the amino acids residues that are performing the modification. The chemical reaction that takes place is often started by a proton transfer to a charged or polar group. It is not uncommon to have a third party involved in the reaction, a cofactor. This can be a simple atom, often a charged metal atom like a zinc atom, or a more complex molecule like the heme group.

Proteins are not the only molecules that can perform tasks in the body. There are also signal substances that mediate signals in and between cells and functional RNA-molecules, for example ribosomal RNA that is the central part of the ribosome, tRNA that transports amino acids for translation as well as other types of non-coding RNA *e.g.* snRNA (involved for example in gene splicing), microRNA and siRNA (involved in gene regulation) [1, 2].

### 1.2.1 Protein regulation

It is, at the moment, estimated to be about 20000 different protein-coding genes in humans [3]. This number is similar to those in other much simpler organisms. The biggest difference between us and for example plants lies instead in the complexity of the regulation. In contrast to prokaryotes, mammals have introns that can be used to regulate gene transcription. Mammals also have several splice sites in most of the genes and many posttranslational modifications that can be used to achieve complex management of protein concentrations, protein activation and degradation. In addition to this, eukaryotes have several different kinds of RNA that are used in additional ways to control the proteins. Furthermore, the DNA of eukaryotes is packed around proteins called histones where the degree of packing is used as a form of regulation.

There are some proteins which are not regulated, but are instead expressed at a constant rate under all circumstances, called constitutive gene expression. Even so, all proteins are degraded after a time to avoid accumulation of high quantities

of proteins which is also a form of control.

### 1.2.2 Genes

The proteins are composed of amino acid residues but the blueprints of the proteins are encoded in our DNA. When more of a certain protein is needed a signal substance is sent to the cell nucleus which upregulates the transcription of this gene. If the protein is involved in a protein complex or a reaction pathway, the signal substance will often affect the transcription of these genes as well. In the transcription process mRNA, or messenger RNA, is created. The mRNA is then transported to the ribosomes where the mRNA is translated into a protein. As there are more amino acid variants than nucleotides in the RNA or DNA three nucleotides are needed to encode one amino acid residue. Each of these triplets, codons, has a corresponding tRNA, transfer RNA, that transports the correct amino acid to the ribosome, see figure 1.1. The ribosome is either located in the cytosol or in the endoplasmic reticulum (ER). If the mRNA sequence has a signal sequence for transportation to ER it will be moved there and usually become a membrane bound protein, otherwise it will be translated by the cytosol ribosomes and used inside the cell.

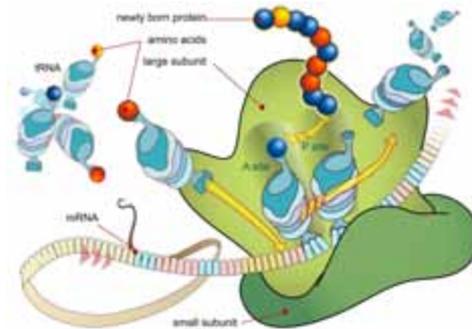


Figure 1.1: The figure shows the translation of mRNA to protein in the ribosomes. Amino acids are provided by tRNA molecules with the codon triplet that corresponds to the next three nucleotides in the mRNA.

## 1.3 Mutations

Mutations are a natural part of evolution but are seldom beneficial for the individual. Most of the mutations will affect DNA outside of the genes and have little effect, other will be silent mutations which do not change the amino acid sequences, while a few will affect the protein function.

Mutations can occur during the cell division process as the DNA in the cell is copied. The DNA can also be damaged by environmental factors, *e.g.* UV-light, free radicals and some human-made chemicals. Most of the DNA damages will be corrected or if that is not possible the cell will usually kill itself. The few mutations that survive can if they affect the DNA repair system or cell replication system increase the risk for cancer. If the mutation effects a germ line cell the mutation can be passed on to the offspring where it can be the cause to one of the numerous diseases currently found in humans.

# Chapter 2

## Methods

### 2.1 Molecular modelling

Molecular modelling is a form of art where the artist and the computer work in cooperation to simulate processes and protein structures that took nature millions of years to evolve. Advanced programs are helpful but in the end it is the input, provided by the user, to these programs that decide the quality of the results. As the calculations are very complex, we need a computer to do the raw number crunching for us. Sometimes, one computer is not enough – we need a cluster or a network of computers.

Examples of what is possible to do using molecular modelling are docking, homology modelling, stability measurements, and active site predictions. These and more examples are further elaborated in the following sections.

#### 2.1.1 Visualization

A commonly used expression is "one picture says more than 1000 words". If the same relationship is true between a picture and an interactive 3D-image then the expression can be modified to "a movable 3D-image says more than 1000 pictures". The essence of this is that by just looking at a protein structure you can gain lots of information and develop theories. A person that is experienced at inspecting protein structures can often make educated guesses that would take the computer a long time, if ever, to figure out. Visualization is also important when you are trying to mediate information to fellow scientists in an effective and understandable fashion. Beside this, the output generated by many of the methods is in the form of coordinates which, if you are not like Cypher in the Matrix, is easiest to comprehend with the help of a visualizer.

### 2.1.2 Protein folding

A protein is composed of a long molecular chain consisting of a number of different amino acid residues (of 20 different types). When this molecular coil is in the correct physiological environment it folds into a specific three dimensional structure. It is in its folded state that the protein performs its primary functions. As many proteins can fold, unfold and then fold again it is only the sequence of amino acids that decide the final folded conformation and the structure should therefore theoretically be possible to determine from the sequence only. This is usually referred to as the Anfisen's dogma [4].

The folding process is driven by the energy difference between the folded and unfolded state. The main problem for the simulation of the folding process is that every amino acid residue can rotate around two single bonds in the main chain. Even if we allow only 3 different angles for each amino acid, representing alpha helix, beta sheet and coil, we get  $3^{100}$  possible conformations for a protein with 100 amino acid residues. Testing all these possible conformations, even using the high rotation rate found in molecules, would take longer than the time the universe has existed. In reality the number of possible conformations is even higher. Despite this, proteins fold in the order of milliseconds to seconds for small single domain proteins. This paradox is called the Levinthal's paradox [5].

The proposed way that nature solves this paradox is by using a guided pathway, often described as a funnel, towards the energy minimum, see figure 2.1. This will not necessarily be the global energy minimum, but rather a very stable local minimum that is easily reached during the folding process. Local interactions and the collapse of the hydrophobic core help to reduce the conformational space. The protein can still get stuck in a number of local minima in the folding process, which will either slow the process or halt it. The latter problem can be solved by chaperone proteins which unfold wrongly folded proteins so that they can start the folding process all over again. In some proteins a specific local minima is almost always present, usually termed molten globule state [6]. Attempts to mimic the natural folding process have so far been mostly unsuccessful [7].

### 2.1.3 Energy minimization

Everything in nature strives to reach a position that is as comfortable as possible, *i.e.* to be in an energy state as low as possible. Proteins are no exceptions to this. This is why the proteins usually fold into a defined structure as this is the lowest energy state given the present environmental factors. Changes in surrounding factors, like temperature, solvent or pH, can change the fold of the protein or make it unfold partially or completely.

To find the global optimal energy of a large protein, given its sequence, is today practically impossible, at least using a computer. As described in 2.1.2, the huge number of conformations possible is almost infinite. An alternative to a brute

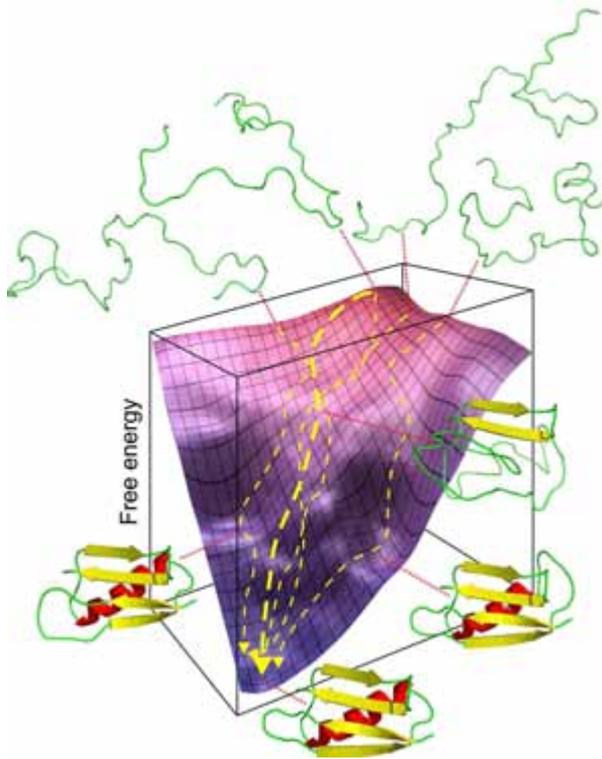


Figure 2.1: The energy landscape of a protein. At high energy the protein is unfolded and at the lowest energy it is in its active configuration. In the folding process several different folding pathways are possible with potential meta stable intermediaries that must be passed.

force method of testing all possible conformations is to use heuristic methods that tries to use smart strategies to search through the energy landscape. The problem is that the protein structure is very sensitive to long range interactions, where the structural context can even shift for instance an alpha helix to a beta sheet which makes it hard to locate the global minimum. Another heuristic approach is to assemble known structural fragments with similar sequence into a complete sequence, see 2.9.2. Programs like ROSETTA can in this way quite often generate the correct general fold of the protein structure.

#### 2.1.4 Monte Carlo based energy minimization

Monte Carlo based methods are often used for finding the lowest energy conformations. The Monte Carlo method is a heuristic technique based on a random walk through the energy landscape. The protein structure is changed locally at a randomly chosen position. While only one residue is chosen in each step, the

surrounding residues are included into the minimization. As the backbone surrounding the chosen residue is free to move, a local change can have a propagating effect on the entire protein. If a lower energy conformation is found the modified structure is kept. Sometimes the structure gets stuck in a local energy minimum, where no locally induced change can improve the energy. To get out of these energy traps there is a certain probability that an unfavourable change is kept. The probability decreases exponentially with increasing energy difference and can be described by eq 2.1,

$$P = e^{\left(\frac{1}{E_{new} - E_{old}}\right)} * T \quad (2.1)$$

where  $E_{old}$  is the total energy of the protein prior to changes,  $E_{new}$  is the total energy after changes are introduced, and  $T$  is the temperature of the simulation system. The temperature is a way to control the probability for unfavourable changes and is increased when no improvements have been observed in a certain number of iterations.

This is the basic version of the method which is not very fast. To speed up the calculations biological knowledge is incorporated into the algorithm in two ways. Firstly, the amino acids which have the highest contribution to the total energy have the highest probability to be in a erroneous conformation. Therefore, the probability is biased towards changing a residue in a high energy environment, and can be described as in eq 2.2,

$$P = \frac{E_{local}}{E_{global}} \quad (2.2)$$

where  $E_{local}$  is the energy of a specific amino acid residue and  $E_{global}$  is the total energy of the protein.

The second speed-up approach exploit the fact that naturally occurring amino acid residues prefer certain angles of their freely rotating covalent bonds, so called rotamers, with transition barriers in-between. Both side chains (except proline side chain) and backbone have rotamers described in a rotamer library, *e.g.* [8]. For example aspartic acid has two chi angles, in addition to the two backbone angles, resulting in nine rotamers. These rotamers are used as starting conformations of the side chain whereupon the local energy environment is minimized.

### Local energy minimization

The easiest way to find the lowest energy in the local environment around a residue is to follow the direction of the most negative gradient as long as this improves the energy. Next, the gradient is recalculated and a step in this new direction is taken and so forth until the length of the step is under a threshold. This method is called gradient descent or steepest descent. Even if this is a robust algorithm it is very slow as it can start zigzagging down energy valleys with flat floors. A much

faster algorithm is the conjugate gradient method [9, 10] which works very well when the energy is already rather close to the energy minimum where the surface is in an approximate quadratic form. To get there, a few initial iterations of the steepest descent are usually used.

The conjugate gradient numerically solves a set of linear equations by an iterative approach. The matrix describing the left hand side of the equation system must be symmetric and positive-definite. When the energy landscape can be described in a quadratic form the method solves the equation system in equally many steps as there are variables. The name of the method comes from the fact that it first takes a step that is orthogonal to the gradient descent, then a step where the direction is almost perpendicular to the first step and finally a step back again, which is similar to the definition of conjugation.

### Simulation length

The suitable number of iterations in the simulations for the global minimization is influenced by the number of degrees of freedom in the protein. The degrees of freedom are determined by the number of possible ways to rotate the backbone and side chains. In the backbone there are the phi and psi angles that can rotate and in the side chain there are the chi angles, of which the average amino acid side chain has two. It is also possible to stop when the minimal energy has not decreased in a pre-set number of iterations, where this number is similarly dependent on the degrees of freedom. In case of folding, the number of iterations needed to get a reasonable structure can be as high as the degrees of freedom to the power of five. For docking, the degrees of freedom are usually limited and for homology modelling the general fold is probably correct from the beginning and the structure can be mainly locally minimized, which decreases the number of needed iterations.

As the method is based on random moves, several simulations of the same system are needed to be able to increase the chances of finding the global optimum. It can also be used to evaluate if the simulation was long enough. If several simulation runs obtain similar energies the result should be of higher quality than if they differ to a large extent.

With the help of Monte Carlo methods it is possible to fold very short peptides with reasonable accuracy. For larger proteins it can be used to investigate small changes or in homology modelling.

#### 2.1.5 Molecular dynamics

Another way to investigate the protein structure is called molecular dynamics. The biggest difference versus energy minimization techniques is that time is introduced, making it possible to study dynamic properties and binding events. The time is not continuous but instead very small time steps are used, usually 1 or 2 fs. The small time step limits the simulation time to the order of milliseconds using modern

computer clusters. There are also more approximate methods where groups of atoms are treated as one particle and more exact methods that use quantum mechanical simulations. The more coarse grained method is faster but not very accurate and the quantum calculations are very time consuming and performed in unrealistic environments such as vacuum and a temperature of zero degrees Kelvin.

## Ensembles

Measurement on a real system will result in properties that are an average of all molecules in that system. In a molecular dynamics simulation only one molecule is studied. However, for a system in equilibrium, the average over time is the same as the statistical ensemble of a multi-molecule system. This means that the properties can be studied in the same way in the simulation as in the test tube.

The position and momentum of an atom in molecular dynamics is usually described in a 6-dimensional space termed phase space. A state in phase space is called a micro-state. Meta-properties like temperature and pressure of a system are called meta-states. For an isolated system in thermodynamic equilibrium all micro-states are equally likely. This means that the macro-state with the highest number of micro-states is the most likely one and that this macro-state has the highest entropy, see eq 2.3

$$S = k_B \log W \quad (2.3)$$

where  $S$  is the entropy,  $k_B$  is the Boltzmann constant and  $W$  is the number of micro-states corresponding to a given macro-state. By the fundamental thermodynamic relation it is possible to deduce the value of useful parameters as for example energy, see eq 2.4

$$dE = TdS - PdV \quad (2.4)$$

where  $E$  is internal energy,  $T$  is absolute temperature,  $S$  is entropy,  $P$  is pressure, and  $V$  is volume.

## Algorithm

In molecular dynamics semi-empirically derived parameters and Newtonian dynamics are used to calculate atom interactions. Forces for individual atoms are integrated over each time step. According to the calculated forces the position and momentum of each atom are updated. This is subsequently done in an iterative manner until a pre-set simulation time is reached. The positions and velocities of the atoms is saved in a trajectory file which can be used for post-simulation analysis.

In the simulation the studied molecule is generally placed in a box full of water molecules. Atoms that get close to the edge of the box will experience abnormal edge effects. To avoid this a periodic boundary condition is used which makes it

possible for atoms close to one edge to interact with atoms close to the opposite edge. Also, atoms going outside one edge will enter the opposite edge.

Most computing time is spent on calculating interactions, so to limit the number of interaction calculations, short range interactions are calculated up to a cut-off distance. For the longer ranged electrostatic interactions this would yield inaccuracies as the potential function only decreases linearly with respect to distance. There are two approaches that solve this with reasonable balance between speed and accuracy: methods based on reaction field or the Particle Mesh Ewald (PME) method [11]. The reaction field approach calculates the interactions up to a cut-off and then assumes a constant dielectric environment for the rest. This works well for homogeneous systems. In PME the interaction beyond the cut-off are instead calculated in reciprocal space as opposed to real space, thereby speeding up the calculations.

### 2.1.6 Energy terms

When calculating the total energy of all interactions in a protein some approximations are needed. The interactions are divided into different categories called energy terms. The parameters for the energy terms are taken from force fields adapted for biological molecules, see 2.1.7. The most important of the energy terms are electrostatic interactions, van der Waals forces, hydrogen bonding, and torsion energy.

#### Electrostatic interactions

Electrostatic interactions are long ranged interactions. This makes them computationally intensive as the number of interactions increases approximately with the cube of the distance. Therefore a number of methods have been developed with focus on either speed or accuracy. The simplest and fastest methods use the Coulombs law with a fixed dielectric constant, described by eq 2.5,

$$F_{el} = \frac{1}{4\pi\epsilon_0} * \frac{q_1q_2}{r^2} \quad (2.5)$$

where  $\epsilon_0$  is the dielectric constant,  $q_1$  and  $q_2$  are the charges of the two interacting atoms and  $r$  is the distance between the two interacting atoms. A more exact but vastly more time consuming method is to numerically solve the Poisson-Boltzmann equation, which is possible to do when implicit solvation is used, *i.e.* when the water molecules are treated as a continuous medium instead of separate molecules.

Strong electrostatic interactions give rise to high specificity between protein and substrate or inhibitor. This can be seen by the relatively high frequency of charged amino acid residues located at active sites and binding pockets.

### van der Waals forces

A much smaller attraction force is the van der Waals force. However, as every atom contributes to this force, the total attraction is still quite large in a protein. The van der Waals force is actually considered by many to be the driving force of folding and the biggest contributor to the stability of the folded protein. The attractive force comes from temporary multipoles in the molecule, called the London dispersion force and is most pronounced in hydrophobic environments of the protein. When the distance between two atoms is too small the force becomes repulsive. The balance between attraction and repulsion can be described by the Lennard-Jones potential which has the form shown in eq 2.6,

$$F_{vw} = 4\epsilon \left[ \frac{\sigma}{r^{12}} - \frac{\sigma}{r^6} \right] \quad (2.6)$$

where  $r$  is the distance between the atom pair,  $\sigma$  is the distance where the potential is zero and  $\epsilon$  is the maximum attraction force [12, 13]. Large and heavy atoms have stronger attraction forces than small and light atoms. The repulsive force increases extremely fast when two atoms get very close to each other. This can be a problem in molecular modelling where one clash can completely dominate the total energy. In ligand screening, where a good binding position should be found very fast, some clashes usually remain after the initial docking. This is because a rigid docking procedure is used, see docking in section 2.2. To solve this, a soft potential can be introduced in order to limit the effect of the clashes. A soft potential is shown in eq 2.7,

$$F_{soft} = \begin{cases} F_{vw} & \text{if } F_{vw} \leq 0, \\ F_{vw} * \frac{t}{t + F_{vw}} & \text{otherwise,} \end{cases} \quad (2.7)$$

where  $t$  is the maximum allowed force. For low repulsive values the forces are almost identical to the original potential function, however the stronger the force the less it contributes to the soft potential function. The attraction force is not affected at all.

### Hydrogen bonding

The hydrogen bond is a special case of dipole–dipole interaction involving a hydrogen atom as acceptor and a heavier atom as donor. The hydrogen bond can be described as a combination of a modified Lennard-Jones potential and an electrostatic calculation of partial charges. The van der Waals part of the potential is described in eq 2.8 where the only difference to eq 2.6 is the attractive part which has an exponent that is 10 instead of 6.

$$F_{hb} = 4\epsilon \left[ \frac{\sigma}{r^{12}} - \frac{\sigma}{r^{10}} \right] \quad (2.8)$$

where  $r$  is the distance between the atom pair,  $\sigma$  is the distance where the potential is zero and  $\epsilon$  is the maximum attraction. The hydrogen bond is direction dependent, so some methods include the angle between donor and acceptor to get a more accurate potential, see eq 2.9

$$F_{hb} = 4\epsilon \left[ \frac{\sigma}{r^{12}} - \frac{\sigma}{r^{10}} \right] \cos^2\theta \cos^4\omega \quad (2.9)$$

where  $\theta$  is the angle between donor and acceptor and  $\omega$  is the angle between acceptor and the direction of the closest lone pair. The electrostatic part is calculated as shown in eq 2.5 using partial charges. Many methods do not use an explicit term for hydrogen bonding. However, if used correctly it can improve the geometry of the atom packing where hydrogen bonding is involved.

### Torsion

The torsion term differs from the terms described above. Instead of intramolecular forces, this term describes the force used to fold the molecule into its present conformation. This is called the dihedral angle deformation force and is described by eq 2.10,

$$F_{to} = K * (1 + sign * \cos(n * T_{angle})) \quad (2.10)$$

where  $K$ ,  $n$ , and  $sign$  (can be either  $-1$  or  $1$ ) are parameters that are depending on the type of dihedral angle and  $T_{angle}$  is the torsion angle.

### Tethers

Sometimes it can be useful to introduce virtual forces that are added to the total energy. This can be done to guide the structure in a certain way. For example if it is known that two cysteines form a disulphide bridge a tether can be introduced between these two residues. The strength of the force can vary, but it is best to use as small force as possible, that will provide the desirable effect, in order to avoid getting stuck in unfavourable conformations. Usually the strength of the force is decreased as the distance approaches the optimal distance. This means that if the protein is in a conformation where the length of the tether is close to optimal the tether will have a very little influence on the total energy. There are two types of tethers; global and local. If the tether is global it effects the structure independent of the length of the tether. The local tethers have a maximum range of effect and can be useful to strengthen interactions once they come close enough. This can be used to guide folding into a certain pathway as the tethers can be adjusted in such a way that they start affecting the structure in a specified order.

## Combination of terms

Depending on what is measured different combinations of energy terms are used. If you want to maximize specificity for a ligand, the most important term is electrostatic interactions, while van der Waals forces are unspecific and largely dependent on the size of the ligand. When studying the effect of mutations on structure and function the hydrogen bond term can be very important as changes in some hydrogen bonds can have large effects on the stability and function of the protein. However, normally all terms mentioned above are used together.

### 2.1.7 Force fields

The force fields used for proteins are often derived from a combination of experiments and quantum level calculations. The force fields describe both bonded and non-bonded interactions. Besides the general functions that describe the interaction potentials the force fields also provide atom specific parameters needed to calculate these potentials. Often several different parameters are needed for each element depending on the surrounding atoms, *e.g.* a carbon in the backbone of a protein or a carbon in a carboxyl group. This makes them approximations of reality and the first level in which errors are introduced.

There are specialised force fields for proteins, like the ECEPP [13] force field used in energy minimization. For molecular dynamics simulations other force fields are used: *e.g.* the GROMOS [14] force field used in GROMACS [15], the AMBER [16] force field used in AMBER, and the widely used CHARMM force fields where CHARMM22 [17] is used for proteins. Many of these force fields are also applied in energy minimizations but are primary for molecular dynamics as they consider all atoms as free variables.

### 2.1.8 CASP

Critical assessment of techniques for protein structure prediction (CASP) is a competition held every second year in order to advance the development of structure prediction programs. The competition is divided into several classes depending on if there are homologous proteins that can be used as templates and if it is a server or manually aided prediction. There are also several categories in which to compete, *i.e.* tertiary structure, residue-residue contact prediction, disordered regions prediction and so on. The structure quality is assessed from a number of different scoring schemes. These are GDT\_TS, ALOP, GDT\_HA, DAL\_4, MAMMOTH and DALI, *cf.* Moulton *et al.* [18]. The most widely utilized of these are GDT\_TS [19] which is used to measure backbone similarity between correct structure and model structure. The score is very effective for template based modelling while it is less useful for new fold predictions where visual inspection of the top scoring models are still needed to choose a winner.

The most recent competition, CASP8, had over 200 research groups participating, so it is a very large happening unlike anything else in the bioinformatics community. However, the competition does not attract as much attention in journals and media as it used to do, so it will be interesting to see how long the competition can stay alive.

## 2.2 Docking

Docking is very similar to energy minimization in the sense that the optimal binding is attained when the total system has the lowest energy. What greatly complicates matters is the total freedom of movement and rotation between the two molecules. Another problem is that the binding might induce large scale changes in order to get the system into the lowest energy state. To predict these large scale changes, without prior knowledge, is almost impossible as it would be as difficult as folding a protein from scratch. Luckily, in most of the cases, only small local structural changes are introduced upon binding.

### 2.2.1 Protein–protein docking

When docking two proteins it is often hard to predict where the binding surfaces of the two proteins are located. The binding energy is usually quite small compared to the size of the molecules and the large binding area. The surface of a protein is mainly hydrophilic to be able to be stable in a water-dominated solvent environment. The binding between proteins therefore must lead to breakage of hydrogen bonds to water molecules and exchange these with their own slightly stronger bonds. One way to find protein binding sites is therefore to look for hydrophobic areas on the surface. If this area is matched with a similar hydrophobic patch on the other protein the complex would potentially be stable.

Protein–protein docking is a very computer intensive task to solve as every point on the first protein must be tested to be docked to the second protein. Also, as the protein is large there is a huge amount of possible conformations for each docking position. To speed up the docking the protein backbone is kept rigid and only the side chains are allowed to move. After this initial docking, the best conformations can be further refined to improve the result.

As protein–protein docking is very hard to model correctly, prior knowledge should be used when possible. Otherwise some kind of experiments should be used to verify the results, *e.g.* mutational replacement of an amino acid residue in the predicted binding site to see if it results in impaired binding.

### 2.2.2 Protein–ligand docking

To dock a ligand to a protein is much easier and also a task that the molecular modelling programs performs with much higher accuracy. It is often possible to find a probable binding site as a functional ligand always binds in a binding pocket. In the case of the active site it is almost always located in the biggest pocket inside the receptor. This greatly limits the number of possible configurations. The docking itself can be performed using a rigid or flexible protein and a flexible ligand. The rigid docking is fast but not especially accurate. If a flexible receptor is used, it is usually only locally flexible around the binding pocket. This generates better docking simulations but takes a longer time.

Even when using a somewhat flexible receptor, ligand–receptor clashes can pose a problem. As atoms get too close together they will repulse each other with an enormously strong force, see eq 2.6. If we instead use the soft potential, described in eq 2.7, we can allow some clashes with a much smaller penalty. This will make it easier to overcome energy barriers to potentially more energetically favourable conformations. It will also decrease the possibility of missing good binding conformations that can be found after refinement. The best conformations from the intital docking calculations are subsequenctly refined where potential clashes are removed and the geometry of the binding pocket is improved.

When setting up a docking simulation it is important to look for coordinated water molecules in the binding pocket. If such molecules exist in the crystal structure they might be necessary for the correct binding of the ligand, either by filling up space or by mediating charge. Another important fact to bear in mind is whether the crystal structure is associated with a ligand or not. If it is not complexed with a ligand the binding pocket can be substantially smaller, or even in a closed conformation. This greatly complicates the docking simulations and decreases the docking accuracy.

### 2.2.3 Virtual ligand screening

Ligand screening can be used to find hitherto unknown substrates or inhibitors to a protein. It is a widely used technique in preclinical trials done by pharmaceutical companies to find lead candidates for further investigation. It is also used in academia, but often on a slightly smaller scale as it is very computer intensive to dock millions of molecules. To speed up the initial screening the receptor is usually treated as rigid, which makes it possible to create a three dimensional grid potential of the receptor. In each point in the 3D map, all active energy terms are calculated and summed up. The closest precalculated value to each atom in the 3D-map can be used to very fast calculate the interaction energy between ligand and the receptor. If the map has high enough resolution the approximation should be of high quality. The most promising ligands can then be refined further with a flexible receptor and a full atom representation, *cf.* Paper VII and [20].

## Ligand library

When doing ligand screening it is theoretically possible to do brute force docking on a huge library. However, it is more efficient to adapt the library based on prior knowledge and practical limitations. An example of prior knowledge can be that the molecule must have a certain kind of functional groups. Practical limitations can be the maximum size that fit in the binding pocket, only testing of non-toxic compounds, and only testing molecules that are commercially available. It can also be useful to remove ligands that are very similar. If a ligand gets a good binding score it can be modified afterwards with small changes to see if the affinity can be increased. A very useful public library of ligands is the PubChem [21] database provided by NCBI that contains over 19 million unique molecules. Here one can search with several limiting criteria such as name, substructure, molecular formula and similarity.

There are other collections of compounds like the Cambridge structural database [22] which is a commercial compound library that contains molecules with determined structures using X-ray and neutron diffraction. Data are collected from open publications as well as direct data deposition and include over a quarter of a million structures. Peptides, oligonucleotides, and inorganic structures are excluded. Another database is the ZINC compound library which is a free database containing 8 million commercially available compounds with 3D-coordinates [23].

## Scoring

Probably the most difficult problem in screening of compounds is how to rank them. There are several problems connected to the scoring. Firstly, large molecules will, in general, obtain a higher binding energy than small molecules due to their larger number of interactions. Secondly, the criteria for good candidate compounds are different depending on if you are looking for a substrate or an inhibitor. Thirdly, as it would take too long time to run extensive refinements on all compounds the scoring must be able to handle clashes between atoms.

The best way to achieve a good rank is to use several different scoring methods and then either take the top ranked from each scoring or make a consensus score for the different scoring methods. If there are known substrates or inhibitors these can be used to obtain weights on the different scoring methods [24, 25]. Methods where the known substrates and inhibitors score high is assigned larger weights while those that score the known substrates low is assigned smaller weights.

The scoring methods can be categorized into three different groups depending on what they base their scoring on. The first group is the force field based methods (*e.g.* Dock [26], Gold [27]) which calculate the binding energy exactly as they are defined in the force field. The second group consists of empirical free energy scoring methods (*e.g.* Chemscore [28], FlexX [29], ICM Score [30]), where some energy terms are assigned greater importance while other can be completely ignored. In

the third group there are methods using knowledge-based potential of mean force (*e.g.* Pmf [25], Drugscore [31]), that use knowledge extracted from the average strength of similar protein–ligand atom interactions in the protein databank (PDB) [32].

In the end, when a top list of candidates have been created, the best way is still to manually examine the compounds and their docking conformations in order to decide which of the compounds to test further *in vitro*.

### Candidate improvement

Once a list of ligand candidates have been created the next logical step would be to try to improve these. The improvements can be done in several ways. Depending on how the original ligand docks, the molecule can be extended or shortened to better fit into the binding pocket or get closer to some interaction partner. Functional groups can also be exchanged to modify polarity, charge or size. From a drug perspective the interaction should be as specific as possible, to avoid side effects from binding to other proteins. Therefore, it is preferred to introduce extra hydrophilic interactions if possible. A less intuitive approach is to decrease the entropy difference upon binding of the ligand. This can be done by restricting the flexibility of the ligand by for example a double bond. The effect is that the difference in freedom of movement ( $\Delta S$ ) become less between bound ligand and free ligand. This will both speed up binding and increase the total binding energy as the ligand is more often in an optimal conformation and will lose less freedom upon binding.

## 2.3 Superimposition

It is often very useful to compare the structural similarity between two proteins in addition to their sequence similarity. This is done by superimposing the two structures upon each other according to some criteria. For evolutionary close proteins the superimposing is done based on a sequence alignment where the distance between aligned residues is minimized in the structure. Otherwise, as the side chains differ so frequently, only the backbone or the alpha carbon is used when comparing the two structures. The problem of finding the optimal structural alignment without a sequence alignment is an NP-complete problem. Therefore, heuristic methods are used instead, *cf.* [33]. The similarity between the two structures is then measured by the root mean square distance (rmsd) between identical matching atoms, see eq 2.11

$$rmsd = \sqrt{\frac{\sum_i^n (a_i - b_i)^2}{n}} \quad (2.11)$$

where  $a_i$  and  $b_i$  are the positions of matching atoms in the two different proteins and  $n$  is the total number of atom pairs. In figure 2.2 two very similar proteins are superimposed with an rmsd value of 1.1 Å.

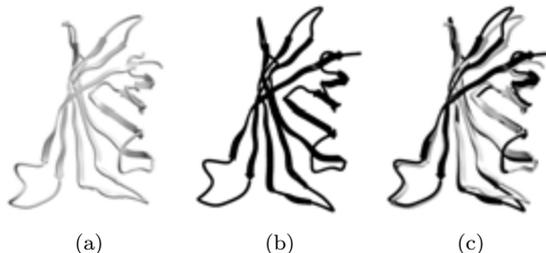


Figure 2.2: Superimposition of (a) and (b) which results in (c)

The figure 2.2 shows an example where the rmsd value is a good measure. However, in some cases it performs less ideal as a measure of similarity. In figure 2.3 the same protein structures are shown, except for the slightly twisted N-terminal in part (b) of figure 2.3. This opens up a small gap in the beta sheet structure. When these two proteins are superimposed as shown in part (c) of figure 2.3 the fit between the structures is slightly out of phase everywhere, leading to an rmsd value of 2.3 Å. However, if the structures were to be locally superimposed on both sides of the slight kink in the structure, both parts gets 1.2 Å in rmsd. The result of the superimposition of the larger part is shown in (d). This demonstrates that one has to be careful when using rmsd values as a measure of similarity.

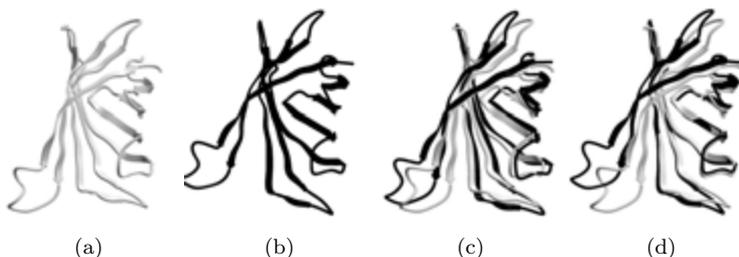


Figure 2.3: Superimposition of (a) and (b) which results in (c). The left part of (b) is slightly twisted which results in a bad superimposition with high rmsd. If instead only amino acid residues 1–99 (out of 133) are superimposed we get (d) with much lower rmsd in that part of the structure.

When superimposing molecules other than proteins it is not possible to use a sequence alignment. One way to solve this is by matching common substructures. If the molecules have no substructure in common one has to look for substructures

that are similar. The resulting rmsd value can be used to evaluate docking results against known bonding conformations or to group similar docking conformations together.

Another measure that is useful is the TM-score [34], especially when comparing structures that have an rmsd value greater than 3 Å. The TM-score gives less penalty to longer distance differences making it less sensitive to large local structural dissimilarities. In this way the similarity of the global topology can be measured. The score can also be used to evaluate a model against a known structure. Values range between zero and one, where a value higher than 0.5 indicates a roughly correct topology.

## 2.4 Homology modelling

Most proteins do not have a solved three-dimensional structure. However, many have a homologue with a solved structure, corresponding to at least part of the protein. The database of homology-derived secondary structure of proteins (HSSP) [35] contains 24% of all human proteins found in Swissprot [36, 37] and provides multiple sequence alignments and structure templates to these proteins. Combined with known structures in PDB 41% of all human proteins in SwissProt have an associated structure. However, the criteria used in HSSP are quite strict in order not to get any false homologues. When studying a specific protein, there often exists more information about the protein than just the sequence. The criteria for homology can then be relaxed and additional homologous proteins can be found and used for modelling purposes.

If a related structure is known it is possible to make a model structure based on this structure using homology modelling. Based on a sequence alignment, the model sequence and the template structure are attached to each other via tethers. A tether is attached between each of the matching amino acid residues in the alignment even if they are not identical. When the lengths of these tethers are minimized the model sequence attains a structure that is very similar to the template structure, actually too similar. There will also be lots of clashes and too few favourable interactions. Monte Carlo energy minimization is then used to remove clashes and optimize interactions, see 2.1.3. The energy minimization is done both globally and locally, where the local minimizations are focused on the loops as they are the most flexible parts. The loops are also more likely to have differences in sequence between proteins with frequent deletions and insertions in addition to a lower degree of conservation, while the fold of the core secondary structures is more conserved.

### 2.4.1 Preparation

The first challenge of homology modelling is to find the best possible template. When there exists several candidates it is usually best to make a model of each and then try to evaluate which one is the best afterwards based on structural quality and energy strain in the model structure.

In order to obtain a high quality structure it is crucial to have a good alignment between the model sequence and the template sequence. To improve the alignment we need to include additional information. This information we can get from the template secondary structure, secondary structure predictions, multiple sequence alignments, conservation analysis and so on. In figure 2.4 we see two different sequence alignments between a part of CYP11B1 and CYP2R. The CYP11B1 is the model sequence and the CYP2R is the template sequence. The sequence identity over the part that has its 3D structure determined is 23%. The relatively low sequence identity makes it very likely that some corrections are needed in the original sequence alignment. The first sequence alignment, figure 2.4 (a), is an ordinary sequence alignment, and the second, figure 2.4 (b), is the corrected sequence alignment based on the template secondary structure. There is a gap of 5 amino acids in the middle of the template alpha helix. If we would make a homology structure from this alignment it would be a very large change in the general structure. The helix would be much shorter and all amino acids in the N-terminal part of the helix would be rotated almost 180 degrees. It might have been acceptable if the gap was moved to the end of the helix but it is more likely that the gap should be in the loop as in the second sequence alignment shown in figure 2.4 (b).

In general, regular secondary structure elements are more structurally conserved than loops making them a fixed reference around which it is possible to build the homology model.

```

Consensus      P      G      TT L      A PN Q      E
CYP11B1      LSPDA-IKANSMELTAGSV----DTTVFPLLMTLPELARNPNVQQALRQESLAAAASIS
CYP2R        NDPSTSPSKENLIFSVGELIIAGTETTTNVLRWAILFMALYPNIQQQVQKEIDLIMGPNG
CYP2R        HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

```

(a)

```

Consensus      S      EL      TT L      A PN Q      E
CYP11B1      -----LSPDAIKANSMELTAGSVDTTVFPLLMTLPELARNPNVQQALRQESLAAAASIS
CYP2R        NDPSTSPSKENLIFSVGELIIAGTETTTNVLRWAILFMALYPNIQQQVQKEIDLIMGPNG
CYP2R        HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

```

(b)

Figure 2.4: Two different sequence alignments between a selected part of CYP11B1 and CYP2R. The first alignment (a) is the original sequence alignment and the second (b) is corrected according to the known CYP2R secondary structure.

Based only on a sequence alignment between two sequences it is impossible to know which amino acid residues that are really important for structure and function. This makes it hard to give priority to which identical residues should be aligned. A multiple sequence alignment (MSA) can solve this issue. The MSA can for example consist of 10 sequences, with 5 related sequences to respectively CYP11B1 and CYP2R. From the MSA it is possible to extract the information about importance of each amino acid residue by relating it to the conservation. This will improve the alignment so that a conserved residue in CYP11B1 more often corresponds to an identical residue that is conserved in CYP2R. If the sequences are at least moderately evolutionary related the resulting alignment usually has less gaps which is good for the quality of the homology model. In figure 2.5 we see another part of CYP11B1 and CYP2R. The alignment in (a) comes from a pairwise sequence alignment while the (b) alignment originates from an MSA. We see that the pairwise alignment has more aligned identical residues (16 vs 11) but with many more gaps. From a structural perspective the MSA-based alignment looks better as there is no gap in regions with regular secondary structure elements and it has a shorter loop that is easier to model.

```

Consensus  G   D   I       NL FG ER                F H E F V L   F
CYP11B1    GSLTLDVQPSIFHYTIEASNLA LFGERLGLVGHSPSSASLNFLHALEVMFKSTVQL-----MFMP
CYP2R      GR-PDFKQLITNAVSNITNLIIFGERF-----TYEDTDFQHMIE-LFSENVELAASAVFLY
Sec. str.  _-__ HHHHHHHHHHHHHHHHHHHHH _------__ HHHHHHHHHH-HHHHHHHHHH_HHHHHHH

```

(a)

```

Consensus          D   I       NL FG ER                F   E       L
CYP11B1           GSLTLDVQPSIFHYTIEASNLA LFGERLGLVGHSPSSASLNFLHALEVMFKSTVQLMFMP
CYP2R             KGRPPDFKQLITNAVSNITNLIIFGERFTYEDTDFQHMIE LFS ENVELAASAVFLYNAF
Conservation      .:      : . . . :      *.*:      .      : .      *
Sec. str.         _____ HHHHHHHHHHHHHHHHHHHHH _____ HHHHHHHHHHHHHHHHHHHHH_HHHHHHHHH

```

(b)

Figure 2.5: Two different sequence alignments between a selected part of CYP11B1 and CYP2R. The first alignment (a) is the original sequence alignment and the second (b) is from an MSA where 5 closely homologous sequences have been used for each of the CYP11B1 and CYP2R sequences.

## 2.4.2 Evaluation

When the homology model is finished it must be evaluated. During this process it is not uncommon that some problems with the structure are found. For example a loop that is too short, causing large strain on the structure leading to unnatural side chain angles and high energy. In this case one must go back and adjust the alignment accordingly and repeat the homology modelling in an iterative process

until the structure has satisfactory quality. Conserved regions in the sequence are also more structurally conserved and are also easier to align. Thus the largest effort must be put on the non-conserved regions. Luckily, as the more important parts of the structure are more conserved the quality of these parts of the model structure will be higher. Therefore, some parts of the model structure can have defects without affecting for example the binding pocket making it possible to perform docking simulations on the model with sufficient accuracy.

There exist many programs that help evaluating the structure. One of the most used programs is PROCHECK [38, 39], used for example to validate crystal and NMR structures, submitted to the PDB database. PROCHECK does several tests on the structure. One of these tests is to check the phi and psi angles in a Ramachandran plot [40], see figure 2.6. In naturally occurring proteins these angles, for all amino acid residues except glycine, are within rather strict boundaries. If a high percentage of the amino acid residues in the model structure is outside of these favourable regions there is probably something wrong. To have some amino acid residues outside is accepted as they can be of special importance where these unnatural angles are crucial for binding or function. Additional checks performed are identification of close contact atoms, non-standard bond lengths and angles in the main chain, and non-planarity of aromatic rings (Phe, Tyr, Trp, His) and end-groups (Arg, Asn, Asp, Gln, Glu).

Other widely used structural validation programs are WHAT IF [41], Verify3D [42, 43], ProsaII [44, 45] and MolProbity [46, 47]. WHAT IF is, as PROCHECK, used to validate structures submitted to the PDB database. It performs similar validation tests as PROCHECK. The WHAT IF program also performs homology modelling and helpful structure related tasks.

Verify3D has a different approach, where a 3D-profile is created of the model which can then be scored in comparison with known structures. By using a sliding window, parts of the sequences that are of low quality can be identified.

In ProsaII both a local score and a overall score of the model quality are obtained. The overall score is compared with PDB structures of similar length to be able to judge if the model has the correct fold. The score is based on a mean force potential calculated from known structures. The potential is based on the spatial separation of two atoms associated with their particular amino acid. This means that for each kind of interaction, *i.e.*  $C_{\alpha} - C_{\alpha}$  or  $C_{\beta} - C_{\beta}$ , there are 400 different potentials. The energy of all intra-molecular interactions is then summed up and a final score is obtained.

MolProbity is more similar to PROCHECK as it can be used to identify clashes and judge the accuracy of a model structure. The method calculates a clash score (z-score) both globally and for each amino acid that can be used to identify bad regions. Clashes and interactions are found using a very small probe of 0.25Å radius rolled over the van der Waals spheres of all atoms. The clash score is then calculated as the sum of the volume for hydrogen bonds plus the sum of the

surface for all van der Waals interactions minus the volume of the overlapping van der Waals spheres, all weighted using various parameters. MolProbity also checks for  $C_\beta$  backbone angle deviations from the norm as a way of finding wrong folds.

## 2.5 *Ab initio* modelling

When there is no homologous structure that can be used as template, the structure must be created from scratch. Even though it is possible to fold proteins using energy minimization the time needed even for small proteins is too large. Therefore, other strategies are used to create the model structure. One way is to simplify the atom representation by grouping atoms together and another is to assemble potentially homologous fragments found in the PDB together with energy minimizations.

## 2.6 Secondary structure prediction

The secondary structure of a protein is of great importance for the protein structure. When globular proteins fold, a hydrophobic core usually forms very fast. This core consists of secondary structure elements consisting of conserved amino acids. Then the rest of the protein folds around this core, limiting the conformational space which helps speeding up the folding process. The high conservation of the core is one important factor why homology modelling is quite successful.

To assign secondary structures to a protein structure is very easy as it is just to measure the psi and phi angles of the backbone. If they are in predefined intervals the amino acid residues are assigned to the corresponding secondary structure. This can be visualized using a Ramachandran plot, where a beta sheet residue should be close to the upper left corner and an alpha helix somewhere below the middle of the left side, see figure 2.6.

To predict secondary structure from sequence is much more difficult as long range interactions can affect the type of secondary structure element. However, the existing tools like PSIPred [48] and Jpred [49] (which is actually a consensus prediction program of several prediction programs) manage to correctly predict secondary structure elements in over 80% of all cases according to the continuously updated benchmark server EVA [50]. PSIPred is based on a neural network that is trained on positions-specific scoring matrices in addition to numerous variables, of which some are described below.

One fact that can be utilized in secondary structure predictions is that different amino acids have preferences towards different secondary structure elements. For example glutamic acid and alanine prefer beta sheets, valine and isoleucine prefer alpha helices, and glycine and aspartic acid are often found in reverse turns. The

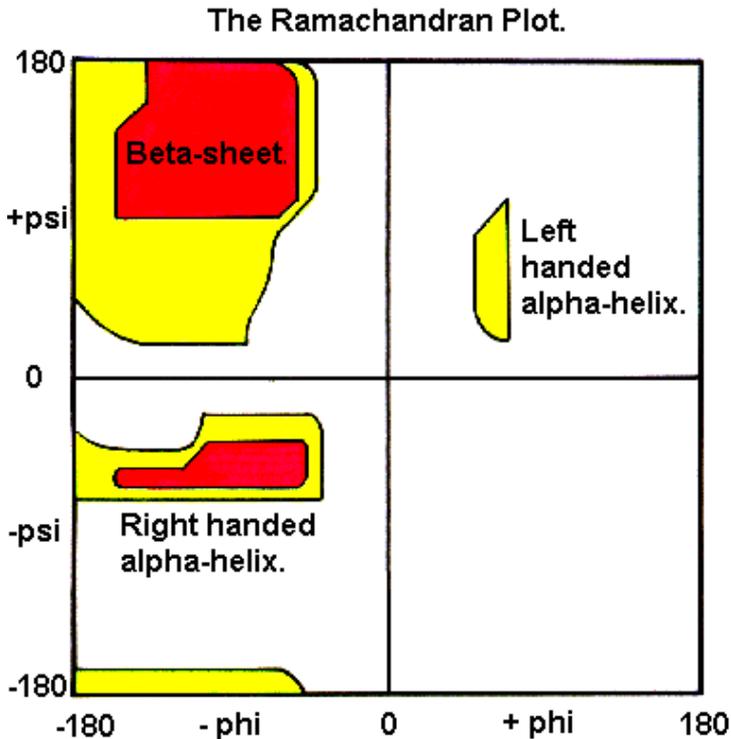


Figure 2.6: A Ramachandran plot where the red areas indicate preferred psi and phi angles for beta sheets and alpha helices. The yellow area are frequented areas of angles for other non-secondary structures in natural proteins. Exception from this is glycine which is more flexible due to its short side chain.

difference in preference for a single amino acid is small, however for a longer stretch of amino acid residues this is often quite informative.

Especially for alpha helices, it is also useful to use the relative position in the secondary structure. As there is a macro-dipole moment inherent in alpha helices the N-terminus is predominated by negatively charged residues and correspondingly the C-terminus is predominated by positively charged residues. Another example is proline which is very seldom found in the middle of an alpha helix as this usually breaks the helix, while it is often found in the first turn. Other amino acids often found in the first turn are for example asparagine and serine, due to the fact that the side chain can form a hydrogen bond to the side chain of the amino acid residue in the next turn of the helix, thereby stabilising the alpha helix. This phenomenon is called N-capping.

Another observation is that the secondary structures often have one side exposed to the surface and one side buried in the core. For beta sheets this means

that the amino acid residues should be alternating between hydrophilic and hydrophobic ones. For alpha helices it means that two or three residues should be hydrophobic and one or two hydrophilic, corresponding to one turn of the alpha helix. Last but not least, conservation analysis is used to incorporate information from homologous sequences, often in the form of profiles.

## 2.7 Methods for prediction

There are many biological questions that can be answered with a prediction based on already known facts. For example, what happens with the protein if a part of the protein is changed? Classification and prediction of data can be useful in both basic research to gain new knowledge and in applied science to help diagnose patients. As an approximation, the answer can often be good enough as a binary answer; yes or no, change or no change, or disease or no disease. The methods in this section are classifiers that make binary predictions. However, many of them can be generalized to make a continuous prediction.

### 2.7.1 Monte Carlo

A Monte Carlo method is a general way of solving an optimization problem using random moves. In Paper II, I used a modified variant of a Monte Carlo method to find the importance of different variables connected to the effect of mutations on protein function. A random variable is chosen and changed to the value that gives rise to the best prediction. This is repeated until a stop criterion is met. Sometimes, the method will get stuck in a local minimum. In this case, a move that decreases the goal function can be used to get out of the minimum.

This method is suitable as long as there are not too many variables and that the variables are linearly correlated with the goal function. If the number of variables is too high the optimization space will be too large to be able to search through in a reasonable time. As only linear relationships can be described by a single weight on each variable, non-linear variables are not correctly modelled. In this case it is better to use an SVM or a decision tree, described in 2.7.3 and 2.7.4.

### 2.7.2 Principal component analysis

Principal component analysis [51], PCA, is a useful mathematical tool that can be used to find patterns in complex data sets with many variables. The input variables, often correlated, are reduced to a few uncorrelated variables, principal components. The first principal component is a vector in the input space where the variability of the data are as large as possible. The second principal component does the same thing for the remaining variability of the data. In this way as much information as possible is captured in very few variables.

As PCA is looking for the highest variability it is important to normalize the input before running the analysis. However, there can still be important variables that are neglected in the first components, because they have low variability in the majority of the data. This can for example be the effect of outliers or that the data are non-linear. The non-linearity can be corrected by a transformation, for example by taking the logarithm of the values. It is also important to remember that PCA only finds linear relationships. This can be mitigated somewhat by making combinations of different variables or taking a higher polynomial of one variable and adding these to the input variables.

The advantage with PCA is that it can find patterns in data without any training data. When training data exists it is often better to use more advanced prediction methods so that this information can be incorporated into the system.

### 2.7.3 Support vector machines

Support vector machines, SVM, are the opposite of PCA in the sense that they increase the dimensions of the input space rather than reduce it as in PCA. The method also needs training data to be able to make a classification. By using a kernel function [52] the input space is transformed into a higher dimensional feature space. In this higher dimensional space a linear classification can be found even though the data are not possible to separate linearly in input space. The data are separated by a hyperplane in feature space. However, this hyperplane can be created in an infinite number of ways. This is solved by choosing data points in feature space, support vectors, that maximize the margin between the two groups and place a hyperplane between these support vectors, see figure 2.7. The size of the margin in the figure is  $2/||\mathbf{w}'||$ , where  $\mathbf{w}'$  is the normal vector from the hyperplane, and  $||\mathbf{w}'||$  is the euclidean length of the vector. By minimizing  $||\mathbf{w}'||$  the margin is maximized [53].

There are many kernel functions that transform the input space into feature space; polynomial, radial basis function, hyperbolic tangent and so on. The polynomial kernel is shown in eq 2.12.

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}')^d \quad (2.12)$$

where  $\mathbf{x}$  and  $\mathbf{x}'$  are two different input vectors, and  $d$  is the polynomial used. The kernel is a matrix with dimensions equal to the input vectors, that is used to transform data between feature space and input space. The kernel function can always be expressed as a dot product in a high dimensional space, making it very easy to calculate even if the function itself is impossible to solve.

The advantages with SVMs are that they can find non-linear separations between classes using linear separation in feature space, making them fast, besides that they are hard to overtrain and thereby predict well on test data. The dis-

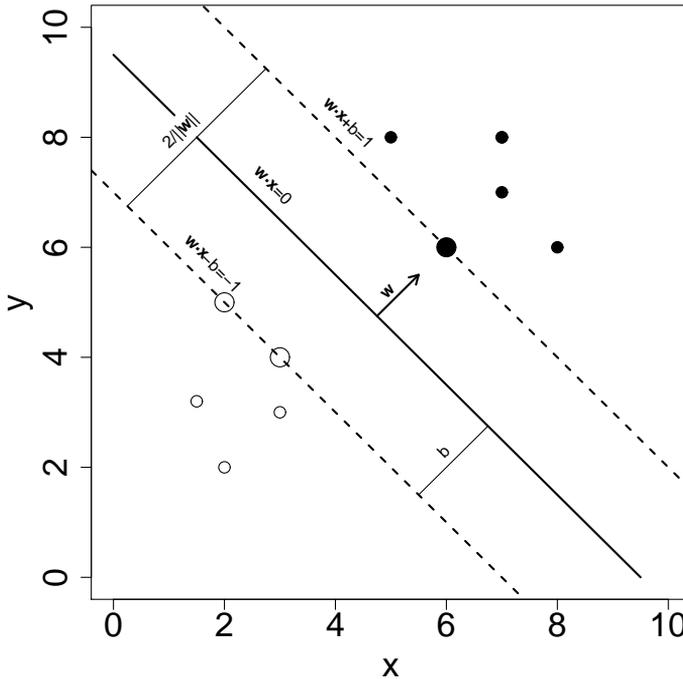


Figure 2.7: An example of a simple classification of data in two dimensions. A linear classification line can be drawn in an infinite number of ways between the two groups. However, by choosing data points, support vectors, that maximize the margin between the two groups the SVM will make better predictions on hitherto unseen data.

advantage is that for many of the popular kernels, the importance of the input variables can not be deduced as the prediction is non-linear.

#### 2.7.4 Decision trees

A decision tree is a rather intuitive way of classifying data where the data are divided into groups, or branches in a tree, at several levels. In every branching a decision is made based on a criterion, most often based on only one variable. A prediction is done when a leaf is reached. The tree can be created automatically or manually, taking advantage of the human experience in the field. Also, the decision tree can be used as a first step where the resulting groups can be further analysed using different classification techniques.

One way to automatically create a decision tree is to find the variable that best split the data according to observations [54]. The same procedure is then repeated for each of the children of the split until no further improvement can be done or

no more splits are possible.

Decision trees capture the fact that the importance of a variable can differ according to the circumstances. In this way a non-linear classifier is created. The drawback is that the method can be overtrained. This can be avoided to some degree by setting a strict stop criterion for where the decision tree should be pruned.

## Random forest

A random forest [55] is an ensemble of decision trees that bases the classification on the most frequent result from the individual decision trees. All the individual trees are fully extended; *i.e.* no stop criterion or pruning. One of two differences between the random forest methods lies in how the branching is implemented. The simplest way is to take a random feature at each branch. The second difference lies in what input data are included when building the tree. Either everything is used, or a random subset of the training data are used. The latter seems to yield better accuracy and less generalization errors.

### 2.7.5 Consensus

When several methods have been applied to the data it is unnecessary to throw away all but the best method. It is better to use them all and let the different methods vote in order to form a consensus. If one method is superior, this method's vote can be weighted higher and *vice versa* for a method that is inferior. In this way several mediocre classifiers can be transformed into a good one, and several good methods into a superior one. This works especially well if the methods work in fundamentally different ways or even better are based on different data. In Paper VI we show how a consensus approach can be used in a powerful way to increase the prediction accuracy on IAPP mutations.

### 2.7.6 Evaluation

As there exists many prediction methods, it is useful to be able to compare how well they perform. A test is usually performed on data not used in the training procedure. The performance can be evaluated in several ways. The simplest way is to take the method that predicts most data correct. However, when data are not evenly distributed this measure can be misleading. A more objective measure is the MCC value described below. It is also useful to find correlation between variables. Sometimes, the predictions can be improved by removing highly correlated variables as this can decrease overfitting, see cross correlation below.

By taking the best method based on the test data, we have in fact done some training on the test data. Therefore, it is valuable to have a third data set which is never used until at the end, where the performance of the method is confirmed. If

enough data exists, this is not a problem, but when data are scarce, the prediction performance can decrease substantially if two different test sets are needed.

### Matthews correlation coefficient

It can be very useful to get a more objective measure of the prediction quality of a two-state classification than percent correct predicted, or accuracy. If the two groups of data are unevenly distributed, a prediction that favours the larger group will get good accuracy, but it can still be a bad prediction. Matthews correlation coefficient (MCC) [56] is such an objective measure of prediction quality. Eq 2.13 shows how the MCC value is calculated.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TP + FP) * (TN + FP) * (TN + FN)}} \quad (2.13)$$

TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative. A perfect prediction will give the value of 1, a random prediction 0, and a perfect negatively correlated prediction a value of  $-1$ .

Very uneven distributions are frequent in bioinformatics, where a common task is to find something specific out of a large sample of data. If we for example are looking for genes associated with a disease, we are expecting to find in the order of 10 genes out of 20000 genes. Even if the FP rate is small, say 1%, and the TP rate is high, say 100%, we would still identify 200 incorrect genes but only 10 correct genes. The MCC value would warn us that this is actually not such a good prediction and give a MCC-value of only 0.18.

### Cross correlation

Similarity between parameters can be measured using the Pearson product-moment correlation coefficient [57] described by eq 2.14.

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \quad (2.14)$$

where  $x$  and  $y$  are values from the two parameters measured, and  $\bar{x}$  and  $\bar{y}$  are the mean values for respective parameter. Values of  $r$  range from  $-1$  to  $1$  where  $1$  means that there is a perfect linear relationship between the two parameters and  $-1$  a perfect negative correlation. Optimal for prediction is if the correlation of the two parameters is close to zero, at the same time as the correlation to the prediction variable is high, as they then complement each other to a high degree. If the real value of the prediction is known the correlation can be used to see which parameters best describe the effect we are looking for and thereby weigh how much each parameter should contribute to the final prediction.

Limitations with this method are that it does only find linear correlations and

that it is sensitive to outliers. The latter is exemplified in figure 2.8 where all examples have a correlation of 0.816. In (a) we see a distribution of data without any outliers and the correlation measure works fine. In (b) we obtain too low correlation because of an outlier while in (c) we have a correlation that is much higher than it should be.

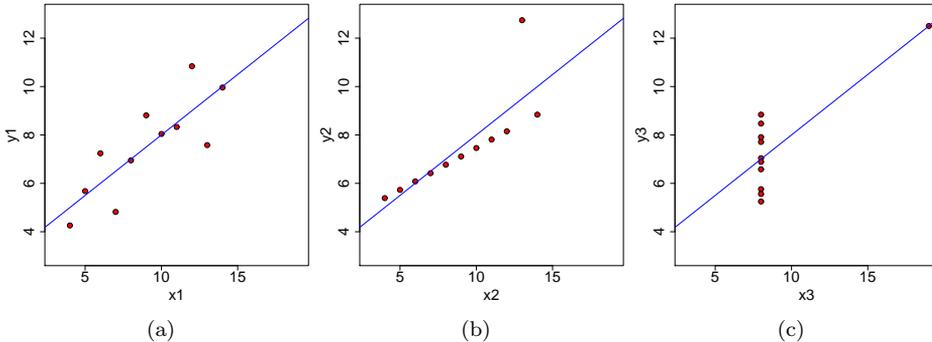


Figure 2.8: Linear correlation between  $x$  and  $y$  for three different data sets with a correlation of 0.816. The data has many common properties; the same mean and variance of the  $x$  and  $y$  variable, the same correlation, and the same regression line.

A method that can be used to automatically remove variables, that have low correlation with the predicted variable, is Lasso [58]. The method minimizes the sum of square errors using linear regression. In addition, Lasso constraints the sum of the absolute values of the parameter weights. The algorithm starts with zero weights for all variables and increments the weights for the variable with the highest remaining correlation to the predicted variable up until the constraint is met or until all parameters have non-zero weight. This means that, for low constraint values, some parameters will get zero weights. By varying the constraint from zero to the infinity, the best linear regression is found. Unnecessary parameters are as a consequence removed entirely.

## 2.8 Protein structure databases

The largest database where structures are stored is PDB [32] containing 59939 structures at the time of writing. 52000 of these are X-ray solved structures while the rest are NMR structures. However, of the 59939 structures only 34549 are completely unique and if 90% sequence identity is used for clustering, 23770 proteins clusters are obtained. Fold classification of these structures is managed in the SCOP (Structural Classification of Proteins) database [59]. The classification is hierarchical where the top level is divided in structural classes depending on the

secondary structure, *e.g.* alpha-helical domains, alpha/beta domains with beta-alpha-beta motifs. There are also a few other classifications for proteins that do not fit into these secondary structure classifications. Next level in the hierarchy is fold followed by superfamily and family. No new folds have been discovered in the last two years which means that the majority of the naturally occurring fold-space for globular proteins seems to be covered by the structures in PDB. A competitor to SCOP is CATH [60, 61] which in contrast to SCOP makes most classifications in an automated fashion. CATH have slightly different hierarchical levels, where the top level is class followed by architecture, topology and homologous superfamily.

## 2.9 Tools

In bioinformatics there are plenty of different tools that perform a wide range of tasks. Often, each task can be solved using several tools that function in subtly different ways. In this section some of the more widely used tools concerning my thesis are presented.

### 2.9.1 Molecular modelling

The molecular modelling program that has been used most extensively in this thesis is ICM from Molsoft Inc., La Jolla, California, USA [62, 63, 64] which is a general purpose molecular modelling program that can perform Monte Carlo based modelling, docking and even include machine learning tools. Other general purpose programs are Chimera [65], CHARMM [66, 67] and Boss (Biochemical and Organic Simulation System) [68].

There are also program packages that work together to perform simulations and analyses, *e.g.* AMBER [69] which can do both molecular dynamics and Monte Carlo modelling and GROMACS [15] which only performs molecular dynamics. There are also more simpler programs for visualization like the popular open source program PyMOL written in Python and the old classical viewing program RasMol which has been around since the early 90's.

### 2.9.2 Homology modelling

In addition to the general purpose programs there are specialised programs like Modeller [70, 71, 72, 73] that makes homology models. However, in my thesis projects, I have used ICM also for this task, partly because it makes good homology models and partly because it makes the integration with other analyses easier. A faster and more easy way to acquire a homology model is to use an on-line server where you just send in a sequence and if there exists a structure with high enough sequence identity, you will get a model structure back within a few minutes. Two of these servers are SWISS-MODEL [74] and ESyPred3D [75]. A more advanced

server is the WHAT IF [41] server, where the user needs to supply their own template and alignment. This requires some more effort but the quality is in the end usually higher. In this server there are also lots of additional useful tools, like structure validation and docking. However, the quality of these models is often low.

### 2.9.3 *Ab initio* modelling

When no homologous structure exists there are programs that try to make *ab initio* or *de novo* models based mostly on the sequence information. ROSETTA is an example of such a program, which also exists in a server version called ROBETTA [76, 77, 78] found at <http://robetta.bakerlab.org/> and as a distributed computer project called ROSETTA@Home found at <http://boinc.bakerlab.org/rosetta/>. ROSETTA uses a fragment based method where fragments are folded based on similar fragments in the PDB database found by PSI-BLAST. In parallel to the original sequence, two additional divergent homologous sequences are folded. When a fragment based method is used a large amount of hypothetical structures is created, called decoys. Probably the hardest part is then to score these models in order to separate the high quality models from the bad ones. One way to remove unnatural structures is to use Pokefind [79] and Knotfind [80] which finds structural features that very seldom occur in nature.

Some *ab initio* folding methods can incorporate additional information from other predictions such as contact predictions, secondary structure predictions and surface accessibility predictions, which improves the chance of getting the general fold approximately correct.

A server for protein predictions that have scored very well in the latest CASP experiments is I-TASSER [81], which can be found at <http://zhang.bioinformatics.ku.edu/I-TASSER/>. The server performs several steps to come up with a number of good candidate structures. First it runs LOMETS [82] which collects target-template alignments from 9 different threading programs. Fragments from PDB are then assembled and energy minimized using a replica exchange Monte Carlo simulation. Unaligned regions, usually loops, are built using *ab initio* modelling. Low energy structures are then identified by SPICKER [83] and run through another energy minimization. The best decoys are chosen by REMO [84] through optimization of the hydrogen-bonding network.

### 2.9.4 Docking

There are specialised programs that perform docking simulations and virtual screening using different amounts of flexibility in the receptor and the ligand. GOLD [85, 86] is one such docking program which uses Chemscore [28] method to find high scoring compounds. GOLD uses a genetic docking algorithm that is opti-

mized for parallel computer environments and virtual screening. GOLD makes it possible to introduce flexibility in the receptor in two ways. The first way is to introduce a soft van der Waals potential, see 2.7, and the second is to specify individual side chains that are allowed to rotate in the active site.

Another virtual screening program is AutoDock [87] in which it is possible to use several different algorithms for finding the lowest energy of a system. These include Monte Carlo simulated annealing, a genetic algorithm, and a hybrid local search genetic algorithm, also known as the Lamarckian genetic algorithm (LGA). In general, the LGA performs the best. The program allows flexible side chains in the protein. The energy is calculated from molecular terms including van der Waals non-bonded interactions, directional hydrogen bonding, screened Coulombic electrostatics, and an atomic solvation parameter-based desolvation free energy term in addition to an empirical term that estimates the loss of torsional entropy of the ligand upon binding.

The degree of success of the docking is largely dependent on the structure quality. Therefore, docking simulations performed on model structures or structures without bound ligands will be less accurate than those performed on a known 3D-structure with bound ligand. Special docking programs have been developed to improve the quality of docking simulations on model and low resolution structures. One of these is Q-Dock [88], which uses a knowledge based pocket-specific potential derived from weakly homologous proteins with bound ligand.

### 2.9.5 Statistical analysis

General purpose statistical programs are very important for analysis of large data sets. Examples of extensively used programs are R, SIMCA-P [89] and Matlab from The MathWorks, Inc. These can for example be used to make predictions, find trends in data and reduce noise in data.

### 2.9.6 Mutation evaluation servers

There are several public servers that perform analysis of the effects of amino acid substitutions on the function of the protein.

#### **SIFT**

SIFT (Sorting intolerant from tolerant) [90, 91] can be found at <http://sift.jcvi.org/>. SIFT uses conservation to judge if a mutation will affect the function of the protein. If a position is conserved most substitutions will be predicted as affecting the function while a non-conserved position will for most substitutions be predicted to be neutral.

## PANTHER

PANTHER [92, 93] can be found at <http://www.pantherdb.org/>. The method is based on position-specific amino acid probabilities in an HMM used to judge mutant severity. The corresponding HMM for a given sequence is collected from PANTHER/LIB that is a collection of HMM:s and MSA:s for protein families.

## PolyPhen

The PolyPhen (Polymorphism phenotyping) [94, 95, 96] server is available at <http://genetics.bwh.harvard.edu/pph/>. The server makes use of several sources of information. It collects functionally important features from SWALL (Swiss-Prot and TrEMBL) [97] and transmembrane predictions from TMHMM [98, 99]. The conservation of each position is determined by a position specific profile with logarithmic ratios of the likelihood for the mutations and the amino acid residue background frequency. If the sequence can be mapped to a known or homologous structure with more than 50% sequence identity, the surface accessibility and changes thereof in addition to side chain volume differences are included in the prediction. The prediction is then performed using empirically derived hierarchical rules.

## CUPSAT

The CUPSAT (Cologne University Protein Stability Analysis Tool) [100] server is located at <http://cupsat.tu-bs.de/>. This server only predicts changes in the stability of the protein, which is often correlated with the function and can be used as a complement to the other servers.

CUPSAT analyses the environment around the substitution by calculating several potentials. The change in potentials between native and replaced amino acid residues is used to make a verdict on the change in stability. The most important potentials are the atom potentials precalculated from the PDB structure for atom pairs between 40 different atoms and torsion angle potentials similarly precalculated for the main chain angles of the 20 natural amino acid residues. The resulting energy calculation is used to classify the mutations. Different cut-off values are used depending on secondary structure and surface accessibility.



# Chapter 3

## Studied proteins

### 3.1 Introduction

Proteins are the entities in our body that actually perform most of the tasks needed to be done in order for us to function. When some of these essential biomolecules become damaged in some way we will function in a non-optimal way. For some proteins there are backup systems and symptoms are very mild while for others the consequences can be devastating. The damage can be in the protein or in the regulation of the protein. In this thesis only the effect of changes at the protein level are studied.

Proteins are really quite fragile, with the difference between a folded functional protein and an unfolded non-functional protein ( $\Delta G$ ) in the order of 5–15 kcal/mol [101]. This is at optimal temperature, pH, and solvent. If any of these factors changed the protein becomes less stable and will, if the environment becomes hostile enough, unfold. This means that the structure is very sensitive to changes, *i.e.* mutations. It is also affected by binding of ligands, interaction with other proteins, and formation of dimers or other multimers. When a ligand binds the structure stabilizes, but as a side-effect there can sometimes be considerable conformational changes. Some proteins need to interact with other proteins to become functional, and still others need to form dimers in order to be stable enough to fold. All these phenomena exist for a functional reason. However, all interactions are not positive for the protein. Sometimes proteins can misfold and aggregate into very stable conformations, amyloid fibres, which then in turn induce more misfolding. This can as a consequence lead to some of our most common diseases such as diabetes and Alzheimer's disease. All this without actually changing anything in the amino acid sequences.

## 3.2 Motivation

The proteins I have studied have been chosen for the fact that there is much information known about these proteins that can be used to gain even more knowledge. Even though extensive research has been performed in connection to these proteins, there are also many unanswered questions left. Some have structural information missing, some miss known connections between function and disease and some lack information regarding protein interactions. The protein studies done in this thesis try to fill some of these gaps in knowledge.

The proteins CYP21, CYP11B1 and p53 are studied from a mutational perspective, where we examine how the mutations affect the stability and function of the protein. In the case of GDNF, NCAM and ANTR1 we are more interested in the actual structure and how they interact with other proteins. We are also looking for functional clues and binding pockets. IAPP is studied from a folding, or rather misfolding, perspective, where we try to find out why and how the peptide forms amyloid fibres and in ADH3 we try to find new substrates and inhibitors by virtual screening.

## 3.3 Steroid 21-hydroxylase

Human 21-hydroxylase (CYP21) is a cytochrome P450 protein that produces cortisol and aldosterone. These hormones are converted from cholesterol in the adrenal glands situated atop the kidneys. 21-hydroxylase deficiency is an inherited disorder also called congenital adrenal hyperplasia (CAH) [102]. The disorder is classified into three different severity classes; salt-wasting (SW), simply virilizing (SV), and non-classical (NC). Aldosterone, of which deficiency is the cause of the most severe class (SW), regulates the salt and water balance in the body, which affects the blood pressure. If the level of aldosterone drops too low it will kill the patient. However, it can be treated with artificial glucocorticoids and salt. When only the cortisol levels are reduced, the patients suffer a milder form of the disease (SV), which causes abnormal growth of the genitalia. For persons where CYP21 is affected but neither aldosterone nor cortisol levels are reduced, a mild form of the disease can still be present (NC), but is usually not noticed until adulthood. The affected can get infertility and various androgen effects. The more severe deficiency classes are found in 1 out of 12000 newborns and the non-classical group is estimated to be as large as 1% of the population. The high frequency of the disease is caused by recombinations with the pseudo-gene CYP21A, which is located very close in the genome.

All proteins in the cytochrome P450 superfamily have a heme co-factor and a typical P450 fold. CYP21 is no exception to this. To perform its function CYP21 needs a redox partner, the cytochrome P450 NADPH oxidoreductase, and a cofactor in the form of a heme group. The enzymatic function of CYP21 is to

add a hydroxyl group to carbon 21 on the steroid, see Figure 3.1.

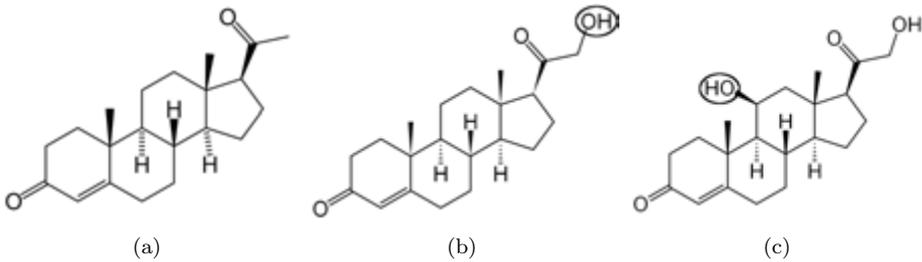


Figure 3.1: Reaction pathway from progesterone (a) to deoxycorticosteroid (b) catalysed by CYP21 and then to corticosterone (c) catalysed by CYP11B1. The added hydroxyl groups in each step are highlighted with a circle.

### 3.4 Steroid 11 $\beta$ -hydroxylase

Steroid 11 $\beta$ -hydroxylase (CYP11B1) is the next enzyme after CYP21 in the cortisol and aldosterone enzyme chain. Patients with CYP11B1 deficiency therefore suffer from the same disease, CAH, as CYP21 deficient patients. However, CYP11B1 is a much less common cause of the disease. The enzyme adds a hydroxyl group to carbon 11 in ring C of the steroid, see Figure 3.1.

### 3.5 p53

p53 or protein 53 is a very important tumour suppressor protein. Mutations that lead to impaired function of p53 are believed to be responsible for about 50% of all cancers in humans [103]. A normally functioning p53 protein will, when DNA-damage is detected, either arrest the cell cycle and hold it in cell state G1/S or induce apoptosis, depending on if the damaged DNA can be repaired. It will also activate DNA-repairing proteins. When the cell cycle is arrested the cell has time to fix the damage without risking that the damaged DNA will propagate by cell division.

Normally p53 is complexed to MDM2, and is thereby inactive. When p53 is activated, by dissociation from MDM2, it functions as a transcription factor for several proteins. It binds to DNA at a DNA-binding site that recognises specific DNA-sequences. Upon binding the protein acts as a promoter for the gene downstream in the DNA sequence. The dissociation is performed by a protein that phosphorylates the amino acid residues in the MDM2 binding region. To regulate the amount of p53 in the cell, the MDM2 gene is up-regulated by p53, and a negative feedback is created.

The protein structure of p53 is unstable, with a low  $\Delta G$  difference between the native and unfolded state [104]. Therefore, it is very sensitive to mutations that affect the structure of the protein. The DNA-binding site is also quite large so there are many additional mutations that affect the function of the protein. Mutant p53 usually accumulates in high concentrations in the cell as the negative feedback of MDM2 is inactivated. This leads to a negative spiral where high concentrations of mutant p53 inhibit native p53 levels. If a cell with a non-functional p53 suffers DNA-damage in a region that regulates cell growth, it can lead to cancer.

p53 consists of five main domains and a total of 393 amino acid residues. The five domains are shown in Table 3.1, of which the central domain is where DNA is bound and also where almost all cancer mutations are found.

Residues	Domain name	Comment
1–42	AD1	Contains the mdm2 binding site and HCD I
40–92	Proline rich domain	Conserved, contains AD2
101–306	Central DNA-binding domain	Contains HCD II to V
307–355	Oligomerisation domain	Tetramerization of p53, contains NES
356–393	C-terminal	Binds to damaged DNA, contains 3 NLS

Table 3.1: The five domains in p53. AD stands for acidic transactivation domain, HCD for highly conserved domain, NES for nuclear export signal and NLS for nuclear localization signal.

### 3.6 Islet Amyloid Polypeptide

Islet Amyloid Polypeptide (IAPP), also called Amylin, is a 37-residue peptide produced in the pancreas by the beta-cells in the islets of Langerhans. It is produced at the same time as insulin, at roughly a hundredth concentration. The physiological function for IAPP is not fully resolved, but recent research shows that IAPP participates together with insulin and glucagon in the regulation of the blood glucose [105].

It is very common for type 2 diabetic patients to have large deposits of IAPP amyloids in the pancreas [106]. The mature amyloid fibril is often described as an inert component without any biological activity even if large deposits can be a physical hindrance for the cell function. Instead, the cell toxic activity is ascribed to the fibril propagation that includes the formation of smaller intermediates often referred to as oligomers or protofibrils [107, 108].

The native peptide of IAPP forms fibres very fast in solution. There is also one known mutation found in humans, S20G, that increases the rate of fibrillation. Furthermore, it has been shown that different parts of the IAPP peptide can form fibres independently [109]. However, there are other animals, like the rat, whose IAPP does not form amyloids even at high concentrations. The sequences differ at six positions, of which three are proline substitutions. It has been shown *in vitro* that proline mutations in human IAPP can hinder fibril formations. For example the S28P and S29P variant does not form any fibres.

### 3.7 Neural Cell Adhesion Molecule

Neural Cell Adhesion Molecule (NCAM) is expressed on the surface of neurons, and additionally on the surface of glia, skeletal muscle and natural killer cells. NCAM is believed to be involved in cell–cell adhesion by inducing neurite outgrowth. The protein has been shown to be of importance for learning and memory.

The protein consists of 7 extracellular domains out of which 5 are Ig-like and 2 are fibronectin type III domains (FNIII) and one additional cytoplasmic domain. Ig-like domains 1–3 has been crystallized and the coordinates are found in PDB with PDB-id 1QZ1. Dimer interaction is suggested to be in either *cis*- or *trans*-formation. In Paper IV we present a model where only the *cis*-formation is possible, and where the dimerization is mediated by the Ig1 and Ig2 domains. The proposed role for the two FNIII domains is that they are involved in signaling.

Regulation of the NCAM gene and protein function is a very complicated process. There exist 27 alternative splicing sites [110]. However, only three isoforms are common, and they vary only in the cytoplasmic domain. There also exist insertions of extra exons of varying length into the NCAM transcript. The VASE domain (VARIABLE domain Spliced EXon) is believed to downregulate neurite outgrowth [111] and the MSD domain (Muscle Specific Domain) is thought to have a positive role in myoblast fusion [112]. To make the regulation even more complicated, the protein can be posttranslationally modified by the addition of polysialic acid (PSA) to the fifth Ig domain. If this modification is removed long-term potentiation (LTP) and long-term depression (LDP) can be abolished [113, 114, 115].

### 3.8 Glial cell derived neurotrophic factor

Glial cell derived neurotrophic factor (GDNF) is another protein active in the brain. The function is not completely known, however some observations have been made. For example it has been shown to potentially promote the survival of certain types of neurons [116]. Especially dopaminergic neurons and motor neurons are promoted, which will wither away in patients having Parkinson's disease and amyotrophic lateral sclerosis (ALS), respectively. The protein seems to be very

important as it is a highly conserved neurotrophic factor.

The structure of rat GDNF has been solved and can be found in PDB with the PDB-id 1AGQ [117] and exists as a homodimer. It has been shown to be a member of a protein complex involving GDNF, NCAM, and GFR $\alpha$ 1 (GDNF family receptor alpha 1), see figure 3.2.

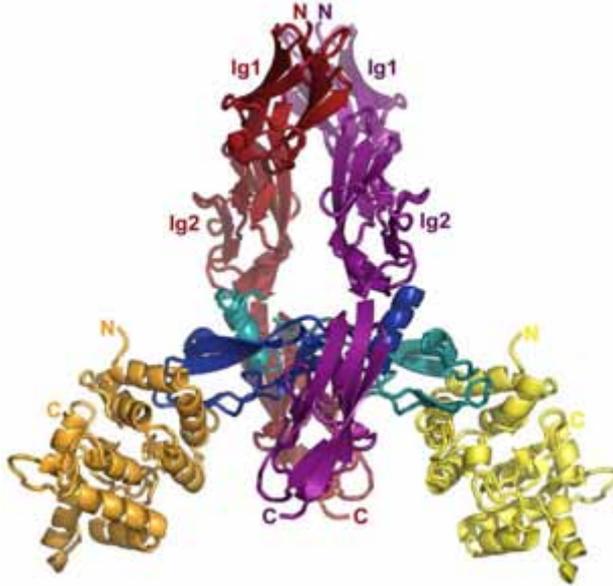


Figure 3.2: The protein complex of the GDNF dimer (blue), the first three Ig-like domains of NCAM (purple) bound to each of the GDNF molecules and GFR $\alpha$ 1 (yellow) also bound to GDNF.

### 3.9 The anion transporter 1

The anion transporter 1 (ANTR1) is an inorganic phosphate transporter located in the thylakoid membrane. The protein is a member of the major facilitator superfamily (MFS) and the SLC17/type I anion transporter family [118]. The protein is found in plants and is highly active in the photosynthesis where inorganic phosphate ( $P_i$ ) is needed to form ATP from ADP. When expressed in *E. coli* the transportation of  $P_i$  has been shown to be dependent on  $Na^+$  while in yeast the transportation is dependent on pH. As opposed to other proteins of the SLC17/type I anion transporter family, ANTR1 is specific for  $P_i$  [119, 120].

### 3.10 Alcohol dehydrogenase class III

Alcohol dehydrogenase class III (ADH3 also known as ADHX or ADH5 or ADH class 3) is the ancestral form in the alcohol dehydrogenase family. The ADH protein family is a member of the MDR superfamily (medium-chain dehydrogenase/reductase). In contrast to the other ADH classes, class III ADH proteins are highly conserved during evolution [121, 122]. ADH3 exhibits high activity towards longer alcohol chains, while for ethanol, the activity is very low. The enzyme is also involved in the degradation of toxic formaldehyde that is a by-product from the ethanol metabolism. Other functions involving ADH3 include NO homeostasis, contribution to retinoic acid formation and oxidation of  $\omega$ -hydroxy fatty acids.



# Chapter 4

## Summary of papers

In this chapter the papers in the thesis are described briefly. The ordering of the papers is such that one paper will in a natural way continue into the next paper and therefore not in chronological order. There are also very brief references to papers not included in the thesis, with me as co-author, as this will enhance the connection between the papers and give some useful background. The focus is on my contributions to the articles while the work of co-authors will only be mentioned in short when it is needed.

### 4.1 Paper I

The focus in my initial minor projects was on structures and SNPs in general. Therefore, it was natural to start a new large scale project concerning mutations that affect protein function and structure. As steroid 21-hydroxylase (CYP21) has over 60 known mutants found in humans, it was used to develop a platform for evaluation of the mutant clinical severity. Through connections by my supervisor, Professor Anna Wedell at Karolinska University Hospital, Stockholm, was contacted and a collaboration was initiated.

Patients with CYP21 deficiency are suffer from congenital adrenal hyperplasia (CAH) which is characterized by impaired cortisol secretion, increased feedback secretion of adrenocorticotrophic hormone (ACTH), and subsequent hyperplasia of the adrenal cortex. In CYP21 deficiency, the levels of aldosterone are often also affected. The severity of the disease is classified into three phenotype classes; salt wasting (SW), simply virilizing (SV), and nonclassical (NC). SW is the most severe form and causes a life-threatening salt loss during the neonatal period. Both SV and SW cause prenatal virilization of external genitalia in affected females. The mildest form, NC, is usually not detected until adulthood, when hyperandrogenic symptoms occur.

The aim of the project was to explain why and how the mutants caused different phenotypes in human patients. To be able to do this we needed to know the structure, which was not available. Therefore, we searched for the closest possible homologue and found rabbit cytochrome P450 2C5 with 31% sequence identity. From this template structure a homology model was calculated. To get a complete model of the CYP21 structure a heme group and a steroid would also need to be added. The template contained a heme group so the position of the template heme group was used as the initial position in the CYP21 structure. As the fold of central parts of CYP450 proteins is very conserved this was judged as an accurate placement of the heme group. To remove clashes between side chains and the heme group the complex was locally energy minimized. The substrate progesterone was subsequently docked into the resulting structure, into a binding pocket next to the heme group. This model was subsequently used to analyse the known mutations.

In Paper SI the H107Y mutant of human carbonic anhydrase II was studied in great detail in collaboration with IFM Biochemistry, Linköping University, where my contribution was to analyse the effect of the mutation *in silico*. Here, we saw that an important network of hydrogen bonds at the active site was disturbed. In a similar manner as for HCAII, the CYP21 mutants were structurally analysed. First they were manually inspected, where structural features and interactions were identified. Then we measured amino acid residue accessibility, residue conservation, hydrophobicity changes, and distance to heme and steroid. Furthermore, we substituted the native residue with the replaced residue and then made global and local Monte Carlo minimizations of the structure. The absolute energy of the mutated structure was evaluated. A total judgment of the measurements and a manual investigation of the structure could then be compared with known clinical phenotypes. The severity class for all but one out of the 60 mutations could be explained using our manual prediction strategy. Based on what was learned from this training set of mutations, new mutations with at that time unknown clinical properties could be predicted into phenotype classes with high correlation to the later determined clinical severity.

Correlations of individual measurements with clinical severity were also investigated. We showed that the energy calculations have clear separations both between normal activity mutations and SV, and between SV and SW. There is also a high correlation between high sequence conservation and severity, and between low residue accessibility and severity.

An additional analysis was performed of the surface of the structure in order to find potential binding sites for protein–protein interaction. CYP21 is known to have a redox-partner interaction, where experimental evidence indicates that basic residues are important as they can interact with acidic residues of the redox-partner. Six basic residues are structurally close together forming a plane that was predicted to be the redox-partner interface. In addition several basic residues were found in CYP21 that when mutated cause CAH.

Apart from the potential redox-partner interaction one more potential binding interface was found. This binding surface is characterized by a hydrophobic area exposed to the surface, consisting of amino acid residues 30–42, 63–66, and 211–219. In this area 85% of the residues are hydrophobic which strongly suggests some functional importance.

In Paper III we set up a server for CYP11B1 where the mutations can be visualized and their predicted severity is shown. At the same time we made predictions on CYP21 using the same methodology which is provided in a similar fashion.

After the CYP21 paper was published we were contacted by two groups of experimental scientists who wanted help with the analysis of their studied proteins using structural bioinformatics. First was a group from Iran at the division of Human Genetics, Immunology Research Centre, Bu-Ali Research Institute. They had found three novel CYP21 mutations that they wanted investigated. The molecular modelling analysis predicted that two of the mutations should cause the SW type of CAH while one was more ambiguous and were judged as SV/SW. The prediction turned out to be fairly correct with all three mutations being of the SW phenotype, *cf.* Paper SII.

The second group, from the Department of Biomedicine and Surgery, Linköping University, was studying protein 17-beta-hydroxysteroid dehydrogenase type 2 (HSD17B2) which is member of the short-chain dehydrogenase/reductase protein family (SDR). They had found evidence of a new polymorphism that could be related to the risk of sporadic and hereditary breast cancer as 17HSD proteins in general are associated with breast cancer. As in CYP21 a homology model was created, using 3-alpha-hydroxysteroid dehydrogenase (PDB-id 1NFQ) as a template, to be able to study the M226V polymorphism. The mutation is situated in the outer region of the active site but it is not strictly conserved, even though its hydrophobicity is conserved between species. However, as valine is also hydrophobic and present in that position in other species the mutation was judged not to affect the function. This was subsequently confirmed by comparing the risk for breast cancer between patients having the M226V polymorphism and patients without the mutation, where no statistically proven difference in risk was found.

## 4.2 Paper II

In Paper I we had found several variables that could be used to help make judgments on the severity of a mutation. However, each mutation needed manual input and time-consuming analysis. In Paper II we decided to investigate the effect of mutations in p53 in collaboration with professor Thierry Soussi, Department of Oncology-Pathology, Cancer Center Karolinska (CCK), KI, Stockholm. p53 is a vital protein that, when DNA damage is detected, will arrest the cell cycle and start DNA reparation. If the damage is too substantial p53 will instead initiate

cell apoptosis. Without a functioning p53 the cells will eventually develop cancer. Actually, disturbances of the p53 function is believed to be involved in 50% of all cases of cancer.

In p53 there are thousands of known mutations found in human cancer patients so the task of evaluating all mutations manually was really daunting. Therefore, we decided to develop an automated method for severity prediction. To do this we needed additional parameters to be able to obtain as high prediction accuracy as possible. We also needed training data for the prediction method. The latter was already available from a group in Japan that had mutated every possible single nucleotide substitution that gave rise to a change in amino acid residue and then measured the activity of the modified proteins in relation to different promoters.

The prediction variables used in Paper I on CYP21 were reused and some new ones were added. In addition to measuring the calculated stability change of a specific residue substitution we measured the average energy of the position by performing simulations on all possible residue replacements. This was used to evaluate the importance of a position in the structure rather than a specific exchange. We also supplemented the change in amino acid properties after mutation with additional parameters, *e.g.* size difference and polarity change according to ProtScale, and amino acid property conservation according to the definition in the ClustalX program.

By using the activity data as training examples, with 25% activity as a separation between severity classes, and with a total of twelve prediction variables, an automated prediction was created. Different approaches were tested, *i.e.* PCA, SVM, PLS, and an in-house developed method. The method developed, PRED-MUT, is based on a Monte Carlo approach where the goal is to find the optimal prediction percentage of both classes. Each variable is assigned a starting weight, whereupon changes in weights are introduced. Changes that improve or do not change the prediction are kept until all possible changes in weights make the prediction worse. Then a random change is introduced to enable escape from the local minimum. The weights of the prediction variables can then be used to make predictions on unknown or test data and to evaluate the importance of each variable.

The resulting prediction method manages to predict the 1148 possible residue exchanges with an overall accuracy of 77%. For non-severe mutations we achieved 74% prediction accuracy and for severe predictions 79% which corresponds to an MCC value of 0.52. Similar MCC values were obtained using SVM and slightly worse with PLS. However, both methods give rise to a more uneven accuracy between classes. The SVM method had also the drawback that the importance of each variable could not be extracted. A subset of cancer mutations found in breast cancer was also evaluated resulting in a prediction accuracy of 88%. The results for each individual mutation are provided via a web server. Here, a 3D-view of the protein is shown with the mutation in question highlighted in the structure.

The prediction variables with the highest individual weights are conservation and accessibility followed by the two energy variables. However, when the weights of the energy parameters were added, the total weight exceeds both conservation and accessibility. The four parameters regarding the amino acid properties were also summed up to a total that is only slightly less than the most important variables. The relatively high information content in the structural energy parameters was probably the reason why the method was able to outperform methods based on sequence only. We also show that both energy variables are individually able to perform as good as or better than the Cologne university based CUPSAT server which calculates stability changes.

### 4.3 Paper III

A protein in the same hormone pathway as CYP21 is human steroid 11-beta-hydroxylase, CYP11B1. Deficiency of CYP11B1 will, in the same way as CYP21, cause CAH. However, the CYP11B1 version of CAH is not equally common. Despite this, there are several known disease related mutations and also a number of probable polymorphisms. As we now had both knowledge of the similar protein CYP21 and a developed protein prediction scheme, we wanted in Paper III to test if this prediction scheme, developed for p53, could successfully be applied on a completely different protein. To acquire all available mutation data we contacted Professor Anna Wedell at Karolinska Institutet/Karolinska University Hospital, Stockholm as she is the current moderator of the CYP11B1 mutation database. In total, 39 missense mutations affecting the CYP11B1 gene were found.

The disadvantage with the original p53 severity scoring in Paper II is that it needs training data. We had too few training examples to obtain reliable weights for all variables, so we decided to use the existing weights found for p53. Instead of using an optimized cut-off, based on training data, we used the median severity score for all mutations to discriminate between severity and non-severe mutations. In the severe class we included mutations causing classical CAH and in the non-severe class we included mutations associated with non-classical phenotype and polymorphisms. In this way the method was able to separate the data with 86% accuracy.

A second prediction scheme, which we called consensus prediction, was also developed to make the prediction less complex and function without training data. To accomplish this we decided to only use the six variables with the highest weights in p53 and to not use the associated weight on these variables. This would work as we only included the most important parameters and these had weights of similar size. Secondly, instead of a continuous severity value, we let the variables vote for if the mutation was considered severe or not. To decide the limit between severe and non-severe we took all values in the protein and ranked them. The

median value would then work as a cut-off. However, this would yield a very sharp border between mutations that could be classed almost equally well in both classes according to a specific parameter. Therefore, we instead made three groups where the middle group was considered as undecided. When a vote was undecided it placed half a vote in both severity classes. To decide the final severity class the votes were counted and the majority decided to which class the mutation should belong. If there are equal numbers of votes for severe and non-severe the mutation was considered to be unclassified.

On CYP11B1 the consensus voting yielded a prediction with an accuracy of 84% without any unclassified mutations. To enable public access to the results a server of all predicted CYP11B1 mutations was developed. Here, the severity of each mutation is presented in addition to a movable 3D-view of the protein where the specific mutation is highlighted. We also used the severity voting to get automated predictions on mutations in CYP21 and in the same way provided a separate server for these.

In order to test if the consensus voting method was applicable in general, we further tested it on 9 evaluation proteins. The 9 proteins are all well studied proteins with a solved crystal structure and at least 20 known mutations with annotated disease phenotypes that could be mapped onto the structure. When we applied the consensus voting prediction on these 9 protein we found that for most proteins the method gave good results with on average 81% accuracy and 8% unclassified mutations. It also performed at least as good as the tested existing methods for all but one of the evaluation proteins.

In certain applications it is more important to obtain a correct classification than to be able to classify all mutations. In this case, the number of mutations without classification can be increased, which will improve the prediction accuracy for the remaining mutations. For example, if we allow 38% of the mutations to be unclassified, the average accuracy will increase to 84% and the MCC value will increase from 0.44 to 0.55. It is also possible to shift the cut-off value in a certain direction in order to improve the accuracy of one of the severity classes. This can be useful if for example it is important to have few severe mutations classified as non-severe as would be the case when doing protein rational design. In this way the prediction algorithm can be adjusted to fit the needs of the application.

## 4.4 Paper IV

This paper is not a continuation of Paper I–III in the sense that it does not involve studies of changes in the protein structure. It does on the other hand involve many of the molecular modelling techniques used earlier in combination with experimental measurements with the aim to study protein interactions and puzzle together a protein complex. This work was done in collaboration with the

Neuroscience group of Carlos Ibáñez at Karolinska Institutet, Stockholm. The primary objective was to study the interaction between the neural cell adhesion molecule (NCAM) and the glial cell line-derived neurotrophic factor (GDNF).

The crystal structure of the GDNF dimer and the first three extracellular domains (out of seven extracellular domains, one membrane domain, and one intracellular domain) of NCAM were known. However, the structure of the interaction interface was not established. From experimental measurements of the interaction between GDNF and NCAM where a single domain was removed in NCAM, it was found that the interaction was located to domain 3 in NCAM. Therefore, NCAM domain 3 was docked to GDNF using protein–protein docking *in silico* in ICM. As the locating of the protein–protein interface was unknown the docking was performed multiple times using starting positions distributed around the surface of GDNF.

In the resulting protein–protein complex four charged residue contacts were identified. Four positively charged residues in GDNF were found within a distance of 2.9 Å to 3.5 Å to four negatively charged residues in NCAM domain 3. Both simulations and experiments show that when the four charged residues in GDNF are mutated to alanine the binding is almost completely abolished. Even though NCAM domain 1 and 2 are very similar in sequence and structure they do not bind to GDNF, thus, to verify the correctness of the predicted protein–protein docking these two domains were also docked. Domain 1 gave rise to substantially less binding energy while domain 2 only yielded slightly lower energy than domain 3. However, as the C-terminal end of domain 2 was pointing into the core of GDNF, this configuration was invalidated if domain 3 was added. The lowest energy structure of domain 2 with a valid configuration had much higher energy than that of domain 3 and thereby the accuracy of the protein–protein docking simulations could be confirmed.

The secondary objective in this study was to add GFR $\alpha$ 1 to the NCAM–GDNF complex. Neither in this case was the interaction site known. Fortunately, there existed a crystallized homologous protein complex containing GFR $\alpha$ 3 and Artemin. The sequence identity between GDNF and Artemin is 37% and between GFR $\alpha$ 1 and GFR $\alpha$ 3 44%. Therefore, it was possible to superimpose the complex onto GDNF and then to superimpose GFR $\alpha$ 1 onto GFR $\alpha$ 3. Even though the exact positioning of the GFR $\alpha$ 1 probably is not perfect the general interaction site should be correct. In this way we could conclude that the interaction between GFR $\alpha$ 1 and GDNF did not clash with the interaction of NCAM and GDNF. As GDNF is a dimer, NCAM binds to both dimers. If the three domains of NCAM, that have known structures, are positioned as found in the protein–protein docking simulations the two NCAM proteins will interact between residues in NCAM domain 1 and 2 without any direct clashes. This further validates the correctness of the predicted protein complex structure.

## 4.5 Paper V

Islet amyloid polypeptide (IAPP) is a short peptide that is secreted in the same cells as insulin, the pancreatic  $\beta$ -cells. The protein, as can be concluded from its name, spontaneously forms amyloids in solution. The formation of IAPP amyloids is strongly correlated with type 2 diabetes. As we had local expertise in the area in the form of Dr. Gunilla Westermark's group at the division of Cell Biology, Diabetes Research Centre, Linköping University, we thought it would be interesting to start a collaboration with them. The aim was to study the amyloid formation of native and mutated IAPP using different molecular modelling techniques. The idea was that studying the amyloid propensity of modified IAPP would generate clues that could help to understand the amyloid formation process.

Variants of the native IAPP peptide in addition to the native peptide were simulated using three different strategies. First, a 12-residue fragment of IAPP (18 to 29), that also forms fibres spontaneously *in vitro*, was investigated using molecular dynamics simulations. Each position was mutated into five different amino acids (cysteine, glutamic acid, leucine, proline, and arginine). The peptides were started in alpha helix conformation and the goal was to measure how long time it took for the helix to break up and how large fraction was found in beta sheet conformation.

The second approach was to do simulations on a larger fragment (residue 11 to 37). As this is more time consuming we restricted the number of peptide variants to the most interesting mutations in the first simulation in addition to previously *in vitro* investigated mutations. This time the peptide was started in a rather stable beta sheet conformation in accordance to models postulated regarding the structure of IAPP located in an amyloid fibre. The aim was to check how the stability of this conformation was affected by the introduction of mutations.

In the third simulation strategy we used Monte Carlo energy minimization to study the interaction between two short IAPP fragments (residue 18 to 29). The peptides were started in beta sheet conformation and energy minimized with an added tether whose function was to get the two peptides in contact with each other. For each simulation run, the starting position for one of the peptides was rotated  $90^\circ$  and so forth until all 10 combinations of starting conformations possible were simulated. Here, we studied the lowest energy of the resulting conformation in addition to the beta content.

Using these three approaches we had a way to predict the amyloid propensity of IAPP variations and native IAPP. The predictions could then be compared with new *in vitro* and previously performed amyloid formation investigations. Each individual simulation approach had an individual prediction accuracy of about 70%. However, as the simulation strategies are fundamentally different they can be successfully combined using a consensus voting. This yielded a prediction accuracy of 94% and an MCC value of 0.88.

Apart from creating an amyloid prediction method we also found several interesting observations concerning the amyloid formation. In general, most substitutions make the peptide less prone to form amyloids. Cysteine and arginine substitutions have the least effect on amyloid formation while proline and glutamic acid substitutions greatly inhibit amyloid formation. In the shorter fragment, the alpha helix starts to unravel from the end, and a core alpha helix, including residues 22 to 25, is still present at the end of many of the simulations. When the beta sheet structure starts to form it is usually residues 21 to 23 and 26 to 28 that are involved, with a beta bend in the middle. This would represent a possible antiparallel stacking of the beta sheets where residue 20 to 23 in one peptide binds to 26 to 28 in the other and *vice versa*. Many substitutions at position 20 and 23 increase or keep the high amyloidogenic propensity of the native IAPP, while residue exchanges at position 24, 25 and 29 greatly decrease the frequency of beta sheets. This indicates that residues 24, 25 and 29 are important for the fibril formation.

All 25 known proteins that today are known to form fibers in humans do so with very similar morphology. Therefore, it should follow that many of the conclusions drawn in the IAPP case should be valid for all amyloid disease causing proteins.

## 4.6 Paper VI

Anion transporter 1 (ANTR1) is a membrane protein found in the thylakoid membrane in *Arabidopsis thaliana*. The protein transports inorganic phosphate. ANTR1 has been previously studied at the division of Molecular Genetics, department of Physics, Chemistry and Biology, IFM, Linköping University and at School of Pure and Applied Natural Sciences, Kalmar University. This work is a continuation of this research where the function and structure of the protein is investigated.

To study the structure of ANTR1 we first made a homology model based on glycerol-3-phosphate/phosphate antiporter (GlpT) from *Escherichia coli*. The template structure with PDB-id 1PW4 had slightly over 20% sequence identity for the residues in the structure. To improve the alignment between template and model, we used a transmembrane helix prediction from the ARAMEMNON database which applies 17 different prediction methods upon the sequence. As the protein was known to have two six-helix bundles, similar to the template, the prediction method that predicted 12 helices, HmmTop\_v2, was used. We also used a multiple sequence alignment (MSA) of ANTR1-6 and SLC17 homologues to improve the modelling procedure.

Three independent homology models were created and the model with the highest quality, based upon a Ramachandran plot, was chosen. The model quality was also evaluated by PROCHECK and was found to be largely correct. In order

to identify functionally important amino acid residues in the ANTR1 sequence, a conservation analysis was performed based on four different MSAs calculated using MUSCLE. The first MSA was obtained using the amino acid sequence of the six members of the ANTR family found in Arabidopsis. The second and third MSAs were obtained from a homology search in the ARAMEMNON plant membrane protein database with criteria of 36% (7 sequences) and 20% (13 sequences) identity, respectively. The fourth MSA was obtained from a BLAST search in the UniProt database (fragments excluded) with an E-value less than  $e^{-40}$  (42 plant and mammalian sequences). By mapping conserved amino acids residues onto the structure we could locate important residues. Five of these were, based on conservation and position, judged to be functionally important for the protein. These five residues (R120, S124, R201, R228 and D382) were subsequently mutated *in vitro*.

To determine the functional role of the exchanged residues, the  $P_i$  uptake was measured in the presence or absence of NaCl. The uptake is normally increased 3–4 times in the presence of NaCl, while for the control this effect is absent. Depending on mutation they either show similar characteristics as the wild type protein, the control or something in-between.  $K_m$  and  $V_{max}$  were also measured in order to judge the effect of involvement in activity and transportation. Finally the ratio  $k_{cat}/K_m$  was calculated in order to be able to compare the transportation effectiveness of the modified with the wild type proteins.

When R120 was mutated to lysine the  $k_{cat}/K_m$  decreased slightly, while a substitution to glutamic acid totally abolished the activity. This indicates that residue 120 is crucial for binding and transportation of  $P_i$ . The lysine mutation shows that a positive charge seems to be the most important property. Similar results were found for R201. The proposed binding residue S124 did indeed affect both  $K_m$  and  $V_{max}$  negatively when mutated to alanine and thereby decreased  $k_{cat}/K_m$ . The only difference between alanine and serine is the missing OH group in alanine. When, on the other hand, the OH group was preserved as in the threonine substitution the uptake of  $P_i$  was even increased. R228 is located on the surface of the protein. Substitution to glutamic acid resulted in loss of uptake activity, while a lysine mutation gave rise to a four-fold lower affinity while  $V_{max}$  was unaffected. This indicates that the residue is important for inorganic phosphate transportation probably by affecting the protein conformational changes upon binding. Another position at the surface of the protein, D382, which is completely conserved in all members of the ANTR and SLC17 family, might be involved in interaction with positive ions due to its negative charge. Mutations to alanine, glutamic acid and asparagine were tested and resulted in increased  $K_m$  and similar or slightly increased  $V_{max}$ . The effect of this is a drastically reduced transportation effectiveness indicating that D382 is, similarly to R228, involved in the protein conformational changes upon binding. One possibility is actually for D382 to form a salt bridge with R228 upon  $P_i$  binding and the following conformational

change needed for completion of the transportation.

## 4.7 Paper VII

Alcohol dehydrogenase class III (ADH3) is a highly conserved protein that exhibits high activity towards longer alcohol chains, while only a very low activity towards ethanol. The enzyme is also involved in the degradation of formaldehyde. Other functions involving ADH3 include NO (nitric oxide) homeostasis, contribution to retinoic acid formation and oxidation of  $\omega$ -hydroxy fatty acids. In the same way as ADH3 has many known functions it also has many known substrates in addition to a number of inhibitors. However, several more ligands should exist. Therefore, we started a virtual screening project of ADH3 in cooperation with Department of Medical Biochemistry and Biophysics at Karolinska Institutet. Almost 41000 compounds were selected from the PubChem database, based on several criteria. To be selected the compounds should include one of the following functional groups: alcohol, aldehyde, carboxy, carboxylic, glutathione or hydroxy. The compound should also have reported bioactivity and have a lower molecular weight than 600 Da.

The selected compounds in addition to 11 known ligands were all docked using ICM in an initial virtual screening step where receptor side chains were treated as rigid. To avoid high energy clashes a soft van der Waals term was used. In the natural reaction process the zinc atom should be close to the active oxygen. Therefore, the distance between the zinc and the closest oxygen in the compound was used to discriminate between binders and non-binders. Only 7% of the compounds have a zinc distance of lower than 2.7 Å compared to 73% of the known substrates and inhibitors. This distance was therefore used to eliminate non-binders. The remaining 2783 compounds were then docked in a second step.

As the number of compounds now was much lower, we could use a more extensive docking simulation on these compounds where we allowed flexible side chains in the receptor. The compounds were subsequently ranked according to five different criteria: binding energy and relevant distances between the compound and the catalytic zinc, NAD<sup>+</sup>, Arg114 and Gln111. Six top scoring compounds that were easily available for purchase were selected for *in vitro* experiments. Two of the compounds were found to have catalytic activity, two to have an inhibitory effect, while the remaining two showed no activity.

In conclusion, by using virtual screening and a scoring function based on known ligands we managed to find four new active compounds out of the 6 top scoring compounds tested.

This work was an extension of Paper SII, a collaboration with IFM Organic Chemistry at Linköping University. In this paper human glutathione transferase A1-1 (hGST A1-1) was reengineered by rational design with the aim to create a

catalyst for thiolester hydrolysis. By doing a single substitution of alanine 216 to histidine the enzymatic function was changed to a thiolester hydrolase.

Different thiolester substrates were screened. In the reaction the thiolester is covalently bound to Y9 so no docking of the substrate was required. Instead the structure of hGST A1-1 was modified first with the A216H mutation and then with the attachment of the thiolester. The distance between the epsilon nitrogen of H216 and the carbonyl carbon of the substrate was measured. The only two of the substrates that had a measured distance below 5 Å were the ones with low  $K_m$  values. Therefore, this distance was concluded to be critical for the reaction velocity and the thiolester selectivity.

# Chapter 5

## Discussion of results

### 5.1 New mutations in CYP21

In Paper I we studied how different mutations in human CAH patients gave rise to different clinical phenotypes. This paper was published in 2006 and since then four new mutations in patients have been found [123]. As all possible mutations are predicted and presented in our CYP21 server it was interesting to compare experimental measurements and clinical phenotypes with the predictions. Table 5.1 presents the mutations with known information and prediction of severity. Clinical severity is shown, where it is known, together with measured activity for 17-hydroxyprogesterone (17OHP) and progesterone. All four mutations exhibit low activity and should therefore be classified as SV or SW. Thus, three out of four mutations show good correlation with our predictions. The three mutations that are correctly predicted, all obtain a high score contribution from the fact that they are highly conserved, buried, destabilizing, and that the change of properties by the amino acid residue exchange is large. Additionally, R426C is close to the heme giving it a maximum severity score. The mutation that is wrongly predicted, C169R, has only a few properties pointing towards it being a severe mutation; *e.g.* causing a large change in property and being in a buried position. Thus, no clear explanation could be found for this mutation.

### 5.2 Mutations

During my PhD studies I have spent a lot of time and effort studying and understanding the effects mutations have on the protein structure and function. The most difficult part is to transform all this knowledge into an automated method that can judge the effect of a mutation as good as or preferably better than an expert in the field can do by manual inspection. When a method makes a predic-

Mutation	Clinical phenotype	Activity in vitro 17OHP / prog.	Prediction	Severity score
C169R	SV	1% / 0%	N / NC	0.3
G178R	N.D.	4% / 0%	SV / SW	0.75
W302R	SW	1% / 0%	SV / SW	0.75
R426C	N.D.	0% / 0%	SW	1.0

Table 5.1: New mutations with known information (clinical phenotype, activity of 17OHP and progesterone) and severity prediction according to our CYP21 prediction server. SW = salt-wasting CAH; SV = simple-virilizing CAH; NC = non-classic CAH; N = normal activity; N.D. = not determined.

tion that is incorrect it is often possible to find an explanation to this by looking deeper into the structure or in the literature. Sometimes this can be used to improve the prediction method but often the findings can be hard to incorporate into the method without a lot of effort and time.

### 5.2.1 Effect on stability

Some proteins are rather stable, making them less sensitive to mutations. However, evolution will always drive the proteins towards being on the brink of unfolding in their native environment. If a protein was super stable, mutations that only affect the stability of the protein would be accepted, and the protein would be less stable. This would go on until the protein would not fold any more and the individual would probably die. This mutation would therefore not be accepted. In this way the evolution has driven the protein to always be on the brink of unfolding. To be somewhat unstable has also the advantage that it makes proteins more dynamic with an improved ability to adapt to substrates and interactions to other proteins.

A destabilized protein will perform its task less fast and an unfolded protein not at all. Thus, influencing the stability of a protein can be as lethal as affecting the function. This is worth remembering as the active site is usually small and unaffected by most mutations. Instability-related mutations are usually found by inspecting residue conservation and inter-residue interactions with nearby amino acid side chains or backbone.

### 5.2.2 Effect on function

Mutations at the active site or close to it will usually impair the function. So if the location of the active site is known these are easy to detect. Then there are other more subtle functional elements that can be affected. If the protein has a co-factor, its binding interface should not be affected and if the protein has interaction partners, their binding should be preserved. The positioning of the co-factor is

usually very important for the enzyme activity, therefore, amino acid changes in the vicinity of the co-factor are often serious. It is more complicated with the binding surface to interaction partners. The amino acid residues in the interaction site can be hard to detect as often unspecific hydrophobic interactions are involved where residues can be exchanged more freely as long as the hydrophobicity is conserved. Also, the interaction could be impaired without affecting the function of the enzyme but there could still be some other function that is disrupted.

### 5.2.3 Measurable variables

In my papers I have used a number of variables to predict mutant severity. In Table 5.2 I have listed most of them. Each of these variables can have a value between 0 and 1, where a value close to 1 should indicate a severe mutation and a value close to 0 should indicate a non-severe mutation. To show that this is true for all prediction variables used in my papers I have listed the difference in average values for each variable between severe and non-severe mutations for all single nucleotide exchanges in p53 in Table 5.2. However, some variables are more evenly distributed between 0 and 1 than others. Therefore, a correction based on the standard deviation is introduced so that the separation ability of the variables can be compared. From the table we can see that conservation and accessibility have the largest separation, followed by the average energy. These are also the same top three variables as in Paper II. However, as some cross correlation between variables exists the weights in Paper II cannot be directly compared with the numbers in Table 5.2. From the table we can observe a standard deviation similar in size as the class difference for the top three variables, while the standard deviation is substantially larger for the rest of the variables. The two classes in p53, defined as below or above 25% activity, are almost equal in size. This implies that the expectation value for all mutations is halfway between the expectation values of the two individual classes. For a variable with a class difference equal to the standard deviation there is consequently half a standard deviation in distance to the total expectation value, which in turn indicates that the variable alone should be able to predict the data with 69% accuracy. Even if this is a good prediction ability for a single variable the accuracy should preferably be higher to be a useful classification. In Paper II we increased the accuracy to 79% by using several variables, three of which had similar prediction ability of almost 70% accuracy. If they were totally independent the accuracy of the combination of the three variables would increase to 77% which shows that our prediction method rather effectively uses the information available in the individual parameters.

#### Stability parameters

An important tool that I have used to predict mutant severity is *in silico* measurements of stability. As described in section 5.2.1 all proteins are sensitive to

Prediction variable	Class difference	Standard deviation	Normalized class difference
Conservation	0.17	0.15	0.26
Accessibility	0.24	0.22	0.26
Average energy	0.10	0.19	0.17
Amino acid similarity	0.19	0.42	0.13
Pocket/Cavity	0.19	0.50	0.12
Individual energy	0.04	0.12	0.10
Secondary structure	0.17	0.50	0.10
Size change	0.07	0.27	0.09
Hydrophobicity change	0.08	0.26	0.09
DNA/Zinc	0.08	0.33	0.08
Polarity change	0.11	0.50	0.07
Similar surroundings	0.02	0.14	0.03

Table 5.2: The difference in average values between severity classes in p53 for different prediction variables. All possible SNPs with over 25% activity are in the non-severe class and the ones below 25% activity are in the severe class. The difference is then normalized based on the standard deviation of the values.

stability changes. The problem is how to measure the effect of a mutation. This we have done using Monte Carlo energy minimizations in combination with local energy minimization procedures. The resulting value is the absolute energy of the protein which can only be used as a relative energy measure compared to other mutations. In this way it is possible to rank all mutations depending on their individual energy. In addition to the individual energy I have also measured the average energy for all possible mutations of a given position. The average energy indicates how sensitive a position is to mutations and the individual energy indicates how large effect a specific mutation will lead to.

The problem with these calculations is that the method focuses on local changes in the structure. However, a mutation that gives rise to a large conformational change will probably also affect the function. Another problem is that the method can give very low energy values for mutations at the active site as these residues often are exposed to the solvent wherein a new side chain has less spatial restraints. Fortunately, active site residues are easily detected by conservation studies, pocket analysis, and by looking for unusual main chain and side chain conformations. The size difference between wild type residue and mutated residue can also have an effect.

Despite these problems, the two energy variables manage to make good predictions, and if their problems could be somewhat amended in a modified version of the parameter they would be even more useful.

## 5.3 Future ideas

### 5.3.1 Additional prediction parameters

During my evaluation of the prediction methods I have found new parameters that could be used to improve the predictions. One idea is to expand the existing similarity of the surrounding parameter from Paper II. This parameter inspects a sphere with 5 Å radius around the investigated amino acid residue and measures the percentage of residues that have the same polarity or charge as the original amino acid. The same sphere could also be checked for other properties; *e.g.* disulphide bridges, salt bridges, and conservation. A disulphide bridge has a well-defined distance (2 Å) between the sulphurs in the two involved cysteine residues. Thus, an examination of a small sphere around each cystein can easily detect probable disulphide bridges. To break a disulphide bridge will have profound effects on the stability of the structure. However, almost only proteins assembled in the ER contain disulphide bridges, in contrast to the reducing environment in the cytosol which breaks the disulphide bonds. Salt bridges are somewhat harder to detect as the force of the electrostatic interaction is decreasing linearly with distance. However, for distances over 5 Å the interaction should be quite small. Both these parameters are rather specialized and will only help the prediction for charged and cysteine residues. Structural conservation, on the other hand, is a parameter that can be used for every amino acid residue. If a non-conserved amino acid residue is located adjacent to many conserved residues, then changing this residue might indirectly influence the structure or function by affecting the surrounding conserved amino acid residues. As the residue itself is not conserved it can obviously be changed to some degree. However, the exchange might be restricted by the surrounding residues in some way, *i.e.* in size or flexibility.

A related parameter to structural conservation is the four-body statistical potential [124, 125, 126]. The natural frequency of groups consisting of four amino acids residues that are closest neighbours have been mapped and compared with the expected random frequency. If the natural frequency is higher than expected by chance the involved residues should, with higher probability, be of importance and therefore have larger effects on the protein structure and function.

In our group we have ongoing studies as well as published works of over- and under-representation of short oligopeptides. In the paper by Bresell *et al.* [127] the over- and under-representation of pentapeptides are investigated. If a pentapeptide is overrepresented it means that its member amino acid residues should be important and substitutions should be avoided. If a mutation changes the pentapeptide into an under-represented one, that change should also be avoided. We have also ongoing studies of finding structural patterns that could be used to identify structural regions of importance.

By searching with the sequence in PROSITE, patterns can be acquired that

describe important features of the protein. If a mutation changes the sequence in such a way that it no longer matches the pattern, the mutations would probably be severe. Many positions, in these kinds of patterns, allow more than one, but not all amino acid residues. This would allow us to find some critical residues that are not completely conserved.

It could also be useful to incorporate annotations of known information about specific positions. For example positions that can be modified by posttranslational changes should be extra sensitive to mutations as it could have several effects, *e.g.* regulation, complex formation, and activation.

In our predictions we use several parameters to determine the change in amino acid property upon residue exchange. Instead, a substitution matrix could be used, *i.e.* the BLOSUM matrix (BLOck SUBstitution Matrix) [128] and the PAM matrix (Point Accepted Mutation) [129]. The advantage of using several parameters is that the importance of each parameter can be adjusted for the specific task at hand instead of having one score for all properties. The risk of using several parameters is that the prediction system can become overtrained. Therefore, the substitution matrix would be useful whenever few training examples exist.

If the phi and psi angles of the backbone of a naturally occurring protein are mapped in a Ramaschandran plot, most residues will fall into certain regions. Residues outside these regions are more likely to be functionally related as their unnatural configurations often cause strain in the structure. This could be used to find functionally related residues that should, if mutated, impair the enzymatic function or binding interactions.

### 5.3.2 Evaluation of some additional parameters

p53 mutations can be used as an evaluation set to test new parameters. To investigate the parameters regarding different properties of amino acid residues in a sphere around a specific residue the p53 structure was used. The most obvious is to use the conservation of the sphere, where the values were taken from the work in Paper II. Firstly, three different measurements was tested, see table 5.3, where the conservation of the mutated residue was included or not, giving a total of six parameters. The inclusion of the mutated residue give rise to a better prediction variable but it is also more correlated with the conservation of the same mutated residue. When it is excluded only new information, not used in the other parameters, is utilized.

The first parameter is calculated from the average of residue conservation for residues that have at least one atom at a distance less than 3 Å from the closest atom in the mutated residue. The second only takes completely conserved residues into account and calculates the frequency of these residues. The third parameter counts the number of residues with completely conserved residues inside the 3 Å sphere. All three parameters have high correlation to the conservation parame-

ter even if the mutated residue is excluded. This is so because if one residue is conserved then many of the surrounding residues are often also conserved. This means that the parameter will only slightly improve the overall prediction. The third parameter is also highly correlated with the accessibility parameter as surface residues will automatically have fewer surrounding residues.

Frequency of different amino acid properties can also be used as parameters. In table 5.3 four parameters using different properties are presented. Firstly, the frequency of charged residues is investigated. A large sphere of 15 Å is used as charged interactions are very long ranged. In the second parameter, polar residues are studied in a 5 Å sphere as interactions are rather local and in the third, frequency of residues that are non-polar are calculated in a 10 Å sphere as non-polar residues are often included in rather large structural features, *e.g.* beta sheets. All these three parameters have reasonable predictability. There is however a high correlation towards accessibility as more polar and charged residues are found on the surface than in the core and vice versa for non-polar residues. In the fourth parameter structural residues were investigated. The amino acids that were counted as structural were proline, glycine and tryptophan. Proline as it bends the backbone, glycine as it can form special structural features and tryptophan for its large size and infrequent occurrence in proteins. However, this parameter was not useful in prediction of the activity in mutated p53.

### 5.3.3 A new prediction strategy

Some of the new and previously used parameters in the severity prediction are only affecting a small fraction of the amino acid residues. Therefore, it could be useful to divide the prediction into a hierarchical prediction, where the specialised parameters are considered first followed by the more general parameters if the specialised parameters do not apply. Furthermore, some of the variables could be nonlinear and behave differently in different situations, *e.g.* buried or on the surface, located in regular secondary or coil structure. This can be captured by a decision tree, or even better a random forest.

### 5.3.4 Consensus

An easy way to improve predictions is to use several methods and then see where they agree and disagree. For example there are several servers that predict the effect of an amino acid substitution, see 2.9.6. Best result will probably be obtained if the methods used are based on different strategies. However, if several methods are very similar, use the best one, to avoid bias towards one prediction methodology.

Prediction variable	Normalized class difference	Cross correlation: conservation	Cross correlation: accessibility
Conservation	0.26	1.00	0.45
Accessibility	0.26	0.45	1.00
Avg. conservation- (3 Å)	0.14	0.42	0.26
Avg. conservation+ (3 Å)	0.14	0.45	0.30
% max conservation- (3 Å)	0.15	0.40	0.24
% max conservation+ (3 Å)	0.20	0.54	0.29
# max conservation- (3 Å)	0.22	0.46	0.54
# max conservation+ (3 Å)	0.25	0.56	0.55
% charged residues (15 Å)	0.17	0.39	0.41
% polar residues (5 Å)	0.17	0.35	0.55
% non-polar residues (10 Å)	0.18	0.36	0.49
% structural residues (15 Å)	0.05	0.11	0.23

Table 5.3: Evaluation of potential parameters compared with previously used ones. A sphere of different radius around a specific amino acid residue is used to find relevant residues. Avg. conservation is the average conservation, % max conservation is the frequency of completely conserved residues, and # max conservation is the number of completely conserved residues. All these variables are measured with and without the mutated residue included (+/-). The % charged residues parameter measures the percentage charged residues inside the sphere, the polar and non-polar residues variables measure percentage polar and non-polar residues, and the structural residues parameter calculates the percentage of structural amino acid residues. All these four parameters include the mutated residue.

### 5.3.5 Future outlook

A more advanced mutation prediction system could in the future, when more crystal structures exist, help in patient diagnosis. When a patient has unclear or unfamiliar symptoms the doctor could send the patient's DNA for sequencing. The sequence could then be uploaded to a mutation prediction server where mutations are found and analysed according to their effect on the protein structure and function. The server will then present which proteins that could be affected, what diseases these could cause and how to cure them. Of course, there need to be a parallel human judgement of the situation.

When more and more structures are crystallized it is also possible find additional connections between them. This is useful when investigating complex diseases that do not depend on one single factor. With the help of new structures it is also possible to construct molecules that could help to stabilize a mutationally unstable protein. This has already been shown to be possible for a few p53

mutations.

Further modelling of amyloid fibres in combination with interaction studies could potentially lead to a way to reverse the fibre formation or inhibit new amyloid formation. This would be a really hard task but extremely rewarding if successful.



# Chapter 6

## Conclusions

In this chapter some of the main conclusions from my time as a PhD student and of my thesis are briefly presented.

- The energy difference induced by mutations can be effectively used to predict mutant severity.
- Prediction of mutant severity can not only be used to infer protein activity but also clinical severity in the form of disease phenotypes.
- Improved predictions can be gained by combining several different methods.
- Even though globular proteins have plenty of differences, the rules that govern their enzyme activity seems to be of similar types.
- Both the virtual screening procedure and the homology models created have been of great use, showing the sophistication of modern molecular modelling techniques.
- Biological systems are very complex and hard to fully understand. One effect can seldom explain everything and sometimes the nature behaves in unintuitive ways. For example, most mutations in IAPP make the peptide less likely to form amyloid fibres. Even so, human IAPP has not evolved in this direction despite the fact that IAPP in other animals has.
- Structural bioinformatics is a very powerful tool that when combined with classical biochemistry can be used to obtain more information in a faster time.
- Bioinformatics is like the spider in the net that interfaces with many other disciplines involved in life sciences. Use this position to your and other scientists advantage to improve both quality and quantity of the work produced.



# Bibliography

- [1] Maciej Szymanski and Jan Barciszewski. Rna regulation in mammals. *Ann N Y Acad Sci*, 1067:461–468, May 2006.
- [2] John S Mattick and Igor V Makunin. Non-coding rna. *Hum Mol Genet*, 15 Spec No 1:R17–R29, Apr 2006.
- [3] Lincoln D Stein. Human genome: end of the beginning. *Nature*, 431(7011):915–916, Oct 2004.
- [4] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A*, 47:1309–1314, Sep 1961.
- [5] C. Levinthal. Are there pathways for protein folding? *Extrait du Journal de Chimie Physique*, 65:44, 1968.
- [6] M. Ohgushi and A. Wada. 'molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett*, 164(1):21–24, Nov 1983.
- [7] M. Karplus. The levinthal paradox: yesterday and today. *Folding and Design*, 2:S69–S75, 1997.
- [8] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*, 6(8):1661–1681, Aug 1997.
- [9] C. M. Reeves and R. Fletcher. Function minimization by conjugate gradients. *Comput. J.*, 7:149–154, 1964.
- [10] E. Polak and G. Ribiere. Note sur la convergence de methodes de directions conjugees. *Fr. Inform. Rech. Operation*, 16:35–43, 1969.
- [11] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An  $w \log(n)$  method for ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.
- [12] J. E. Lennard-Jones. -. *Proc R Soc London A*, 106:463–477, 1924.

- [13] F.A. Momany, R.F. McGuire, A.W. Burgess, and H.A. Scheraga. Energy parameters in polypeptides, vii: Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.*, 79:2361–2380, 1975.
- [14] L.D. Schuler, Daura X., and W.F. van Gunsteren. An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry*, 11:1205–1218, 2001.
- [15] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman J C Berendsen. Gromacs: fast, flexible, and free. *J Comput Chem*, 26(16):1701–1718, Dec 2005.
- [16] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem*, 24(16):1999–2012, Dec 2003.
- [17] Jr. MacKerell, A.D., D. Bashford, M. Bellott, Jr. Dunbrack, R. L., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, III W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorcikiewicz-Kuczera, D. Yin, , and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.
- [18] John Moult, Krzysztof Fidelis, Andriy Kryshchak, Burkhard Rost, Tim Hubbard, and Anna Tramontano. Critical assessment of methods of protein structure prediction-round vii. *Proteins*, 69 Suppl 8:3–9, 2007.
- [19] Adam Zemla. Lga: A method for finding 3d similarities in protein structures. *Nucleic Acids Res*, 31(13):3370–3374, Jul 2003.
- [20] M. Schapira, B. M. Raaka, H. H. Samuels, and R. Abagyan. Rational discovery of novel nuclear hormone receptor antagonists. *Proc Natl Acad Sci U S A*, 97(3):1008–1013, Feb 2000.
- [21] J. Kaiser. Science resources. chemists want nih to curtail database. *Science*, 308(5723):774, May 2005.
- [22] Frank H Allen. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallogr B*, 58(Pt 3 Pt 1):380–388, Jun 2002.

- [23] John J Irwin and Brian K Shoichet. Zinc—a free database of commercially available compounds for virtual screening. *J Chem Inf Model*, 45(1):177–182, 2005.
- [24] C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *J Med Chem*, 43(25):4759–4767, Dec 2000.
- [25] Reiji Teramoto and Hiroaki Fukunishi. Supervised consensus scoring for docking and virtual screening. *J Chem Inf Model*, 47(2):526–534, 2007.
- [26] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J Mol Biol*, 161(2):269–288, Oct 1982.
- [27] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267(3):727–748, Apr 1997.
- [28] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des*, 11(5):425–445, Sep 1997.
- [29] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, Aug 1996.
- [30] M. Totrov and R. Abagyan. Derivation of sensitive discrimination potential for virtual ligand screening. *Annual Conference on Research in Computational Molecular Biology*, 3:312–320, 1999.
- [31] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol*, 295(2):337–356, Jan 2000.
- [32] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.
- [33] Leandro Martínez, Roberto Andreani, and José Mario Martínez. Convergent algorithms for protein structural alignment. *BMC Bioinformatics*, 8:306, 2007.
- [34] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, Dec 2004.

- [35] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68, 1991.
- [36] UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Res*, 35(Database issue):D193–D197, Jan 2007.
- [37] Cathy H Wu, Rolf Apweiler, Amos Bairoch, Darren A Natale, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Raja Mazumder, Claire O'Donovan, Nicole Redaschi, and Baris Suzek. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–D191, Jan 2006.
- [38] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. Stereochemical quality of protein structure coordinates. *Proteins*, 12(4):345–364, Apr 1992.
- [39] R. A. Laskowski, M. W. Macarthur, D. S. Moss, and J. M. Thornton. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.
- [40] G. N. RAMACHANDRAN, C. RAMAKRISHNAN, and V. SASISEKHARAN. Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7:95–99, Jul 1963.
- [41] G. Vriend. What if: a molecular modeling and drug design program. *J Mol Graph*, 8(1):52–6, 29, Mar 1990.
- [42] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, Jul 1991.
- [43] R. Lüthy, J. U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364):83–85, Mar 1992.
- [44] Markus Wiederstein and Manfred J Sippl. Prosa-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*, 35(Web Server issue):W407–W410, Jul 2007.
- [45] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17(4):355–362, Dec 1993.
- [46] Ian W Davis, Andrew Leaver-Fay, Vincent B Chen, Jeremy N Block, Gary J Kapral, Xueyi Wang, Laura W Murray, W. Bryan Arendall, Jack Snoeyink, Jane S Richardson, and David C Richardson. Molprobit: all-atom contacts

- and structure validation for proteins and nucleic acids. *Nucleic Acids Res*, 35(Web Server issue):W375–W383, Jul 2007.
- [47] J. M. Word, S. C. Lovell, T. H. LaBean, H. C. Taylor, M. E. Zalis, B. K. Presley, J. S. Richardson, and D. C. Richardson. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol*, 285(4):1711–1733, Jan 1999.
- [48] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2):195–202, Sep 1999.
- [49] Christian Cole, Jonathan D Barber, and Geoffrey J Barton. The jpred 3 secondary structure prediction server. *Nucleic Acids Res*, 36(Web Server issue):W197–W201, Jul 2008.
- [50] V. A. Eyrich, M. A. Martí-Renom, D. Przybylski, M. S. Madhusudhan, A. Fiser, F. Pazos, A. Valencia, A. Sali, and B. Rost. Eva: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, 17(12):1242–1243, Dec 2001.
- [51] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.
- [52] V. Kecman. Learning and soft computing - support vector machines, neural networks, fuzzy logic systems. *The MIT Press*, 2001.
- [53] B. Schölkopf and A. J. Smola. Learning with kernels. *MIT Press*, 2002.
- [54] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- [55] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [56] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta*, 405(2):442–451, Oct 1975.
- [57] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66, 1988.
- [58] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267–288, 1996.
- [59] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.

- [60] Alison L Cuff, Ian Sillitoe, Tony Lewis, Oliver C Redfern, Richard Garratt, Janet Thornton, and Christine A Orengo. The cath classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*, 37(Database issue):D310–D314, Jan 2009.
- [61] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath—a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, Aug 1997.
- [62] R. Abagyan and M. Totrov. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol*, 235(3):983–1002, Jan 1994.
- [63] R. Abagyan, M. Totrov, and D. Kuznetsov. Icm - a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, 5:488–506, 1994.
- [64] M. Totrov and R. Abagyan. Efficient parallelization of the energy, surface, and derivative calculations for internal coordinate mechanics. *Journal of Computational Chemistry*, 10:1105–1112, 1994.
- [65] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612, Oct 2004.
- [66] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1982.
- [67] Jr. MacKerell, A.D., B. Brooks, III Brooks, C. L., L. Nilsson, B. Roux, Y. Won, and M. Karplus. Charmm: The energy function and its parameterization with an overview of the program. *The Encyclopedia of Computational Chemistry*, 1:271–277, 1998.
- [68] William L Jorgensen and Julian Tirado-Rives. Molecular modeling of organic and biomolecular systems using boss and mcpro. *J Comput Chem*, 26(16):1689–1700, Dec 2005.
- [69] David A Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668–1688, Dec 2005.

- [70] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, Dec 1993.
- [71] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.
- [72] A. Fiser, R. K. Do, and A. Sali. Modeling of loops in protein structures. *Protein Sci*, 9(9):1753–1773, Sep 2000.
- [73] Narayanan Esvar, Ben Webb, Marc A Marti-Renom, M. S. Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using modeller. *Curr Protoc Protein Sci*, Chapter 2:Unit 2.9, Nov 2007.
- [74] T. Schwede, J. Kopp, N. Guex, and M.C. Peitsch. Swiss-model: an automated protein homology-modeling server. *Nucleic Acids Res*, 31:3381–3385, 2003.
- [75] Christophe Lambert, Nadia Léonard, Xavier De Bolle, and Eric Depiereux. Esyspred3d: Prediction of proteins 3d structures. *Bioinformatics*, 18(9):1250–1256, Sep 2002.
- [76] Dylan Chivian, David E Kim, Lars Malmström, Philip Bradley, Timothy Robertson, Paul Murphy, Charles E M Strauss, Richard Bonneau, Carol A Rohl, and David Baker. Automated prediction of casp-5 structures using the rosetta server. *Proteins*, 53 Suppl 6:524–533, 2003.
- [77] David E Kim, Dylan Chivian, and David Baker. Protein structure prediction and analysis using the rosetta server. *Nucleic Acids Res*, 32(Web Server issue):W526–W531, Jul 2004.
- [78] Dylan Chivian, David E Kim, Lars Malmström, Jack Schonbrun, Carol A Rohl, and David Baker. Prediction of casp6 structures using automated rosetta protocols. *Proteins*, 61 Suppl 7:157–166, 2005.
- [79] Firas Khatib, Carol A Rohl, and Kevin Karplus. Pokefind: a novel topological filter for use with protein structure prediction. *Bioinformatics*, 25(12):i281–i288, Jun 2009.
- [80] Firas Khatib, Matthew T Weirauch, and Carol A Rohl. Rapid knot detection and application to protein structure prediction. *Bioinformatics*, 22(14):e252–e259, Jul 2006.
- [81] Yang Zhang. I-tasser server for protein 3d structure prediction. *BMC Bioinformatics*, 9:40, 2008.

- [82] Sitao Wu and Yang Zhang. Lomets: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res*, 35(10):3375–3382, 2007.
- [83] Yang Zhang and Jeffrey Skolnick. Spicker: a clustering approach to identify near-native protein folds. *J Comput Chem*, 25(6):865–871, Apr 2004.
- [84] Yunqi Li and Yang Zhang. Remo: A new protocol to refine full atomic protein models from c-alpha traces by optimizing hydrogen-bonding networks. *Proteins*, 76(3):665–676, Aug 2009.
- [85] G. Jones, P. Willett, and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol*, 245(1):43–53, Jan 1995.
- [86] Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved protein-ligand docking using gold. *Proteins*, 52(4):609–623, Sep 2003.
- [87] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19:1639 – 1662, 1999.
- [88] Michal Brylinski and Jeffrey Skolnick. Q-dock: Low-resolution flexible ligand docking with pocket-specific threading restraints. *J Comput Chem*, 29(10):1574–1588, Jul 2008.
- [89] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold. *Multi- and Megavariable Data Analysis - Principles and Applications*. Umetrics Acad., Sweden, 2001.
- [90] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Res*, 11(5):863–874, May 2001.
- [91] Pauline C Ng and Steven Henikoff. Accounting for human polymorphisms predicted to affect protein function. *Genome Res*, 12(3):436–446, Mar 2002.
- [92] Paul D Thomas, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–2141, Sep 2003.
- [93] Paul D Thomas, Anish Kejariwal, Nan Guo, Huaiyu Mi, Michael J Campbell, Anushya Muruganujan, and Betty Lazareva-Ulitsky. Applications for protein sequence-function evolution data: mrna/protein expression analysis and coding snp scoring tools. *Nucleic Acids Res*, 34(Web Server issue):W645–W650, Jul 2006.

- [94] S. Sunyaev, V. Ramensky, and P. Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*, 16(5):198–200, May 2000.
- [95] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe, A. S. Kondrashov, and P. Bork. Prediction of deleterious human alleles. *Hum Mol Genet*, 10(6):591–597, Mar 2001.
- [96] Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous snps: server and survey. *Nucleic Acids Res*, 30(17):3894–3900, Sep 2002.
- [97] H. Hermjakob, F. Lang, and A. Apweiler. Sptr - a comprehensive, non-redundant and up-to-date view of the protein sequence world. *CCP11 Newsletter*, 2.3, 1998.
- [98] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–182, 1998.
- [99] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580, Jan 2001.
- [100] Vijaya Parthiban, M. Michael Gromiha, and Dietmar Schomburg. Cupsat: prediction of protein stability upon point mutations. *Nucleic Acids Res*, 34(Web Server issue):W239–W242, Jul 2006.
- [101] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, Aug 1990.
- [102] Saroj Nimkarn and Maria I New. Steroid 11beta- hydroxylase deficiency congenital adrenal hyperplasia. *Trends Endocrinol Metab*, 19(3):96–99, Apr 2008.
- [103] T. Soussi and C. Bérourd. Assessing tp53 status in human tumours to evaluate clinical outcome. *Nat Rev Cancer*, 1(3):233–240, Dec 2001.
- [104] Stefan Bell, Christian Klein, Lin Müller, Silke Hansen, and Johannes Buchner. p53 contains large unstructured regions in its native state. *J Mol Biol*, 322(5):917–927, Oct 2002.
- [105] Davida F Kruger, Catherine L Martin, and Christopher E Sadler. New insights into glucose regulation. *Diabetes Educ*, 32(2):221–228, 2006.
- [106] P. Westermark. Quantitative studies on amyloid in the islets of langerhans. *Ups J Med Sci*, 77(2):91–94, 1972.

- [107] Magdalena Anguiano, Richard J Nowak, and Peter T Lansbury. Protofibrillar islet amyloid polypeptide permeabilizes synthetic vesicles by a pore-like mechanism that may be relevant to type ii diabetes. *Biochemistry*, 41(38):11338–11343, Sep 2002.
- [108] Rakez Kaye, Elizabeth Head, Jennifer L Thompson, Theresa M McIntire, Saskia C Milton, Carl W Cotman, and Charles G Glabe. Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science*, 300(5618):486–489, Apr 2003.
- [109] P. Westermark, U. Engström, K. H. Johnson, G. T. Westermark, and C. Betsholtz. Islet amyloid polypeptide: pinpointing amino acid residues linked to amyloid fibril formation. *Proc Natl Acad Sci U S A*, 87(13):5036–5040, Jul 1990.
- [110] A. A. Reyes, S. J. Small, and R. Akeson. At least 27 alternatively spliced forms of the neural cell adhesion molecule mrna are expressed during rat heart development. *Mol Cell Biol*, 11(3):1654–1661, Mar 1991.
- [111] P. Doherty, C. E. Moolenaar, S. V. Ashton, R. J. Michalides, and F. S. Walsh. The vase exon downregulates the neurite growth-promoting activity of ncam 140. *Nature*, 356(6372):791–793, Apr 1992.
- [112] Misa Suzuki, Kiyohiko Angata, Jun Nakayama, and Minoru Fukuda. Polysialic acid and mucin type o-glycans on the neural cell adhesion molecule differentially regulate myoblast fusion. *J Biol Chem*, 278(49):49459–49468, Dec 2003.
- [113] C. G. Becker, A. Artola, R. Gerardy-Schahn, T. Becker, H. Welzl, and M. Schachner. The polysialic acid modification of the neural cell adhesion molecule is involved in spatial learning and hippocampal long-term potentiation. *J Neurosci Res*, 45(2):143–152, Jul 1996.
- [114] Luminita Stoenica, Oleg Senkov, Rita Gerardy-Schahn, Birgit Weinhold, Melitta Schachner, and Alexander Dityatev. In vivo synaptic plasticity in the dentate gyrus of mice deficient in the neural cell adhesion molecule ncam or its polysialic acid. *Eur J Neurosci*, 23(9):2255–2264, May 2006.
- [115] Oleg Senkov, Mu Sun, Birgit Weinhold, Rita Gerardy-Schahn, Melitta Schachner, and Alexander Dityatev. Polysialylated neural cell adhesion molecule is involved in induction of long-term potentiation and memory acquisition and consolidation in a fear-conditioning paradigm. *J Neurosci*, 26(42):10888–109898, Oct 2006.
- [116] G. Martucciello, I. Ceccherini, M. Lerone, and V. Jasonni. Pathogenesis of hirschsprung's disease. *J Pediatr Surg*, 35(7):1017–1025, Jul 2000.

- [117] C. Eigenbrot and N. Gerber. X-ray structure of glial cell-derived neurotrophic factor at 1.9 Å resolution and implications for receptor binding. *Nat Struct Biol*, 4(6):435–438, Jun 1997.
- [118] Christian Roth, Gerhard Menzel, Jean MacDonald-Comber Petétot, Sylvie Rochat-Hacker, and Yves Poirier. Characterization of a protein of the plastid inner envelope having homology to animal inorganic phosphate, chloride and organic-anion transporters. *Planta*, 218(3):406–416, Jan 2004.
- [119] Lorena Ruiz-Pavón and Angel Domínguez. Characterization of the *yarrowia lipolytica* ylsrp72 gene, a component of the yeast signal recognition particle. *Int Microbiol*, 10(4):283–289, Dec 2007.
- [120] B. Guo, Y. Jin, C. Wussler, E. B. Blancaflor, C. M. Motes, and W. K. Versaw. Functional analysis of the arabidopsis pht4 family of intracellular phosphate transporters. *New Phytol*, 177(4):889–898, 2008.
- [121] R. Kaiser, B. Holmquist, B. L. Vallee, and H. Jörnvall. Characteristics of mammalian class iii alcohol dehydrogenases, an enzyme less variable than the traditional liver enzyme of class i. *Biochemistry*, 28(21):8432–8438, Oct 1989.
- [122] B. Persson, J. Hedlund, and H. Jörnvall. Medium- and short-chain dehydrogenase/reductase gene and protein families : the mdx superfamily. *Cell Mol Life Sci*, 65(24):3879–3894, Dec 2008.
- [123] Yulia Grischuk, Petr Rubtsov, Felix G Riepe, Joachim Grötzinger, Svetlana Beljelarskaia, Vladimir Prassolov, Natalya Kalintchenko, Tatyana Semitcheva, Valentina Peterkova, Anatoly Tiulpakov, Wolfgang G Sippell, and Nils Krone. Four novel missense mutations in the *cyp21a2* gene detected in russian patients suffering from the classical form of congenital adrenal hyperplasia: identification, functional characterization, and structural analysis. *J Clin Endocrinol Metab*, 91(12):4976–4980, Dec 2006.
- [124] C. W. Carter, B. C. LeFebvre, S. A. Cammer, A. Tropsha, and M. H. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J Mol Biol*, 311(4):625–638, Aug 2001.
- [125] Alexander Tropsha, Charles W Carter, Stephen Cammer, and Iosif I Vaisman. Simplicial neighborhood analysis of protein packing (snapp): a computational geometry approach to studying proteins. *Methods Enzymol*, 374:509–544, 2003.
- [126] Ewy Mathe, Magali Olivier, Shunsuke Kato, Chikashi Ishioka, Pierre Hainaut, and Sean V Tavtigian. Computational approaches for predicting the

- biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res*, 34(5):1317–1325, 2006.
- [127] Anders Bresell and Bengt Persson. Characterization of oligopeptide patterns in large protein sets. *BMC Genomics*, 8:346, 2007.
- [128] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, Nov 1992.
- [129] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. "A model of evolutionary change in proteins", *Atlas of Protein Sequence and Structure (volume 5, supplement 3 ed.)*. Atlas of Protein Sequence and Structure (Vol 5, Supplement 3). National Biomedical Research Foundation, 1978.