University of Alberta

Computational support systems for prediction and characterization of protein crystallization outcomes

by

Marcin Jerzy Mizianty

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems

Electrical and Computer Engineering

Fall 2013

Edmonton, Alberta

This thesis is dedicated to my wife

for her constant support and unconditional love.

Abstract

Analysis of protein structures may reveal their function, regulation and interactions. Almost 90% of the known protein structures were solved using X-ray crystallography; however, many more structures remain unsolved. Protein Structure Initiative (PSI) project was created to speed up structure determination. PSI includes structural genomics (SG) centers that perform high-throughput crystallization which processes hundreds of proteins using standardized protocols. Large quantities of crystallization data generated by PSI fueled research that looked into proteins' properties associated with success of crystallization. In spite of intense research crystallization of proteins is still among the most complex and least understood problems in structural biology. Since SG centers do not focus on individual proteins, but rather on covering the protein structure space, they have certain flexibility in selection of targets. At the beginning of my PhD program we designed and assessed three accurate methods that predict crystallization propensity based on a protein sequence. These methods could be used to prioritize targets based on their predicted propensity for the successful structure determination. We observed that as the crystallization protocols are updated the predictors of crystallization propensity need to be correspondingly upgraded and enhanced. To this end, in the course of the thesis we developed an accurate predictor that generates crystallization propensity and indicates causes of the potential crystallization failure, which can occur at any of the three major steps in the protein crystallization protocol: production of protein material, purification, and production of crystals. Our predictors are empirically compared against state-of-the-art in the field demonstrating favorable predictive performance. Finally, we designed another accurate and runtime-efficient method which we then used to perform first-of-its-kind large-scale

analysis of crystallization propensity for proteins encoded in 1,953 fully sequenced genomes. Analysis of these predictions shows that current X-ray crystallography combined with homology modeling could provide an average per-proteome structural coverage of 73% with over 60% coverage for archaea and bacterial proteomes, and between 35 and 70% for eukaryotes. Moreover, our study revealed that use of knowledge-based target selection increases coverage by a significant margin, which for majority of organisms is between 25 to 40%.

Acknowledgments

First and foremost, I would like to express my deep gratitude to my supervisor Dr. Lukasz Kurgan for all his guidance, passion, motivation and most of all for the tremendous amount of time he spent to make me a better researcher. I could not have imagined having a better mentor for my PhD studies.

I would like to especially thank my wife Agata for her love, motivation, understanding, and for leaving everything behind so we could be together. You are making me a better person.

I would like to thank my daughter Larysa for bringing so much joy and happiness to my life.

I would like to thank my parents and the rest of my family for their unconditional love and support, their encouragement, and for their understanding of my choices.

I would like to thank my friends for their support, motivation and time spent together.

I would like to thank my fellow lab members for their collaboration and support.

I would like to extend my gratitude to Killam Trusts and University of Alberta for their financial assistance during my PhD studies.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3D – three dimensional

Å – Angstrom

AA(s) – Amino Acid(s)

AUC – Area Under the ROC Curve

CF – Crystallization Failed

Cryo-EM – Cryo-Electron Microscopy

CRYS – Crystallizable

DNA – Deoxyribonucleic acid

FN –False Negatives (positive trials that were predicted as negatives)

FP – False Positives (negative trials that were predicted as positives)

GO – Gene Ontology

GPCR(s) – G- Protein Coupled Receptor(s)

MCC – Matthew's Correlation Coefficient

MF – Material Failed

mRNA – messenger RNA

NMR – Nuclear Magnetic Resonance

PDB – Protein Data Bank

PF – Purification Failed

pI – isoelectric point

PSI – Protein Structure Initiative

RBF – Radial Basis Function

RNA – Ribonucleic acid

ROC – Receiver operating characteristic

SER – Surface Entropy Reduction

SG – Structural Genomics

SVM – Support Vector Machines

TN – True Negatives (correctly predicted negative trials)

TP – True Positives (correctly predicted positive trials)

tRNA – transporting RNA

UK – United Kingdom

USA – The United States of America

Chapter 1

Introduction

Proteins are major components of all living organisms and are involved in virtually every aspect of biological systems. Analysis of their three dimensional structures helps to reveal their function, regulation and interactions (Harrison, 2004; Chang et al., 2013). The most common way to solve these three dimensional structures, which accounts for almost 90% of the known protein structures (source: http://www.rcsb.org/pdb/statistics/holdings.do), is by using X-ray crystallography. The protein crystallography started in early XX century when first protein crystal structure was obtained (Sumner, 1926), and the first X-ray crystallography was attempted (Bernal & Crowfoot, 1934). In late 1950s' the first protein structure was solved (Kendrew et al., 1958), which led to the creation of the Protein Data Bank (PDB) (Berman et al., 2000) in 1971. PDB started with only 7 structures and this number has been growing exponentially ever since (Berman et al., 2012). Currently PDB is the main repository of protein structures, and as for 30th June 2013 it holds 89,189 protein structures. For the first few decades majority of protein structures depositions were made by many, usually small, research groups scattered across the world. These efforts concentrated on individual proteins and lacked centralized control, which would assure that duplication does not occur and that the new depositions are useful for a broader research community. At the same time major genomic projects, such as the human genome project, generated millions of protein sequences and it became clear that effort in solving structures cannot keep up with the pace of growth in sequences. This motivated researchers to develop fast computational approaches to predict protein structures. Arguably, the most promising approaches are based on homology modeling (Ginalski, 2006). They predict structure based on the similarity in sequence with another protein for which a structure is known, i.e. a template protein. The ever-widening gap between the number of known protein sequences and structures and a need for large database of diverse structure templates (to improve homology modeling), motivated in late 1990s' a

more integrated and “protein family” oriented approach for protein structure determination. To this end, a few projects including Protein Structure Initiative (PSI) (Terwilliger et al., 1998) and Structural Genomics Consortium (Williamson, 2000) were initiated. These efforts generated funding for over a dozen Structural Genomics (SG) centers, i.e., centers which aim at determination of proteins’ 3D structure on a large scale. Rather than trying to solve individual proteins, the SG centers that are part of PSI concentrated on selecting representatives from each protein family; this resulted in enlarging the set of templates and potentially served a wider group of end users. Moreover, these centers developed high-throughput crystallization approaches, which decreased the associated costs (Joachimiak, 2009). The high-throughput was achieved by processing a large number of proteins, at the same time, using standardized protocols. These protocols would not be optimized for any given protein, but instead, they perform well for most of proteins, and they simultaneously screen hundreds of crystallization conditions for each protein target. As PSI started generating large quantities of crystallization data, the corresponding information about crystallized targets and applied crystallization protocols was recorded (Stevens, 2000). However, initially the potential benefits of these data could not be fully exploited as they lacked information about the unsuccessful attempts (Rodrigues & Hubbard, 2003), which began to be stored in 2004 (Kouranov et al., 2006). Availability of these data fueled research that looked into biochemical and biophysical properties of proteins that are associated with their propensity for crystallization. Unfortunately, in spite of intense research crystallization is still among the most complex and least understood problems in structural biology (Hui & Edwards, 2003). Only as little as 2 to 10% of pursued protein targets yield high-resolution protein structures (Service, 2005). Furthermore, more than 60% of the cost of structure determination is consumed by the failed attempts (Slabinski et al., 2007a). However, since SG centers do not focus on individual proteins but rather on representatives from protein family, they have certain flexibility in selection of targets. Therefore, accurate methods which could guide crystallographers in this target selection process, i.e., methods which predict crystallization propensity based on a protein sequence, are needed. These methods save resources of the SG centers, which are currently spent on the unsuccessful targets, by helping to choose protein targets for which probability of successful structure determination is higher.

1.1 Thesis statements and goals

Our aim is to develop computational methods that use protein sequences to provide outcomes that support target selection for X-ray crystallography. This is motivated by the following three observations. First, we observed that as the crystallization protocols are updated, the target selection methods need to be correspondingly upgraded and enhanced (Mizianty & Kurgan, 2011). Second, the current developments in target selection area are focused on the prediction of crystallization outcome and do not provide insights into causes of the crystallization failure, which can occur at any of the three major steps in the crystallization protocol: production of protein material, purification, and production of diffraction quality crystals (Mizianty & Kurgan, 2011). Third, methods that predict crystallization propensity should be used to estimate attainable structural coverage which can be obtained using X-ray crystallography (in combination with homology modeling) as this could provide useful insights to plan future directions. To this end, we focus on the problem of accurate prediction of protein crystallization output and its steps and analysis of predicted crystallization propensity on genomic scale. We define the following thesis statements:

- The accuracy of current predictors can be improved;
- Prediction of crystallization propensity requires continuing development and improvement, as advances in crystallization protocols may render previously unsolvable targets to be solvable;
- Methods can be built to support multiple steps in the X-ray crystallography based structure determination pipelines;
- Predicted crystallization propensity can be used to estimate structural coverage that is attainable using current crystallization protocols and homology modeling.

To address the abovementioned thesis statements we define three goals:

1. **To develop methods which provide more accurate prediction of crystallization propensity when compared to existing predictors.** Our first goal concentrates on accurate prediction of the outcome of the (entire) crystallization attempt from a given protein sequence, without providing a more detailed feedback in the case of a negative outcome. To address this goal we designed three crystallization propensity

predictors which output probability of the input protein to be crystallized by an SG center using standard protocols.

2. **To predict outcomes of individual steps in the crystallization protocol.** In our second goal we still focus on the sequence-derived prediction of the crystallization propensity, however in case of the negative predictions (i.e., protein cannot be solved using crystallization) we indicate which step in the crystallization protocol is the most likely cause of the failure. This is a first-of-its attempt to address this goal. The beforehand knowledge of a potential cause of the crystallization failure may help crystallographers to modify a protein in a way that it will be more likely to pass the crucial step. To fulfill this goal, we designed a method that predicts the outcomes of the three major crystallization steps.
3. **To compute and analyze of the attainable structural coverage.** In our final goal, we use the predicted crystallization propensity to perform first-of-its-kind large-scale computation and analysis of the attainable structural coverage using current crystallization protocols and homology modeling. We utilize a snapshot of protein "universe" composed of around 10 million proteins from all fully sequenced proteomes collected in July of 2012. We analyze the structural coverage and the corresponding functional coverage of the protein structure universe that can be attained and analyze differences in the coverage between different superkingdoms of life. This goal required the development of a new crystallization propensity predictor that is sufficiently fast to process such large protein set. The predictive quality of predictors that were developed in the first goal deteriorated with time and the predictor developed in the second goal was not fast enough.

We believe that our methods will find interest at the SG centers and smaller crystallography labs in support of the target selection for X-ray crystallography. Some of the existing methods for prediction of the crystallization outcome are already in use the by SG centers (Slabinski et al., 2007a; Overton & Barton, 2006; Price et al., 2009; Babnigg & Joachimiak, 2010). Our claim is substantiated by our successful collaboration with crystallographer Prof. Joachimiak, who is a PI at the Midwest Center for Structural Genomics. Moreover, new physicochemical characteristics of input protein sequences, which were found during the development of our methods, may help crystallographers

to better understand the process of protein crystallization, which ultimately may lead to improvements in the crystallization pipelines and protocols.

1.2 Outline

Chapter 2 provides background information on proteins, proteins structure, and methods to solve the structure. It also briefly overviews the history of Structural Genomics and summarizes former attempts to discover physicochemical properties of proteins by analysis of the data from first large-scale crystallization projects. Finally, it discusses related works concerning crystallization propensity prediction and gives background about machine learning, which is used to build our predictors, and methods and criteria used to evaluate predictions.

Next, in Chapter 3 we address the first goal and present approaches for the crystallization propensity prediction. We first provide a brief overview of relevant datasets and next we describe each of our three methods separately. Our first method, CRYSTALP2, was designed to mitigate the sequence length constrains and to improve over the first predictor developed in our lab (prior to my studies) CRYSTALP. After designing CRYSTALP2, we analyzed predictive performance of methods that were available at the time and found that their predictions were complementary. To exploit that, we designed a meta-predictor, MetaPPCP, which combined predictions from different existing methods to generate more accurate results. Finally, we utilized novel (in this field) information that was extracted from the input protein sequence to design the third predictor, CrysPred. These three methods, and other existing predictors, are empirically compared and the findings are summarized.

Since low predicted crystallization propensity provides limited amount of information to crystallographers who need to pursuit a given target, in Chapter 4 we describe and assess our approach to provide/predict more information for the unsuccessful crystallization targets. First, we define a new protocol to obtain and annotate (using appropriate database) unsuccessful trials with the step in the crystallization pipeline for which they failed. Next, we describe PPCpred, a first-of-its-kind method which is able to predict crystallization output and, in case of predicted failure, points to the step that is the most likely to be the cause of the failure. Our

method is then empirically compared with existing crystallization propensity predictors and this part of the research is summarized.

Chapter 5 describes our large scale analysis of predicted crystallization propensity for the current snapshot of protein “universe”. First, we describe the data that were used. Since the existing crystallization propensity predictors could not be used to perform this analysis, next we introduce our new predictor. Following that, we overview the attainable structural coverage computed with our method and present the corresponding conclusions.

The last chapter presents our conclusions and summarizes this research. We list major contributions and findings and outline possible future research directions.

Chapter 2

Background and Related Work

2.1 Background on proteins

2.1.1 Proteins

Proteins are (arguably) the most important components of all living organisms. They define various types of living beings from the simplest bacteria and viruses to the most complex mammals, such as humans. They make up to 50% of the dry weight of cells and are involved in virtually every reaction in biological systems. Proteins were recognized as a distinct class of biological molecules in the XVIII century, however the name “protein” was first proposed in 1838 by Jöns Jakob Berzelius. This term is derived from the Greek word πρωτειος (proteios), meaning "of primary importance". Berzelius chose this term as he knew that proteins are important for animal nutrition. Studies of the scientific nature of proteins have started as early as in 1838 when Gerardus Johannes Mulder, following an advice from Berzelius, described the chemical composition of proteins. He hypothesized that proteins must consist of similar blocks which differed only slightly in their chemical compositions. It took over a century to provide the answer to this hypothesis, when in 1949 Frederick Sanger demonstrated that proteins consist of linear polymers of amino acids (AAs) (Sanger, 1949) linked together by peptide bonds (which was one of the theories since 1902). Sanger received the Nobel Prize in Chemistry in 1958 for his work on sequencing of bovine insulin which resulted in this discovery.

Amino acids are the building blocks of proteins and their arrangement in a protein is responsible for protein’s shape and function (Anfinsen, 1973). Each AA consists of amine group, a carboxylic acid group and a side-chain; see Figure 2.1. The peptide bond between neighboring AAs is established between carboxyl and amine groups, therefore enabling AAs to link together, and leaving the groups on both ends open which enable a further growth in a linear chain of AAs. The chain ends with the amine group on one side and the carboxyl group on the other side. These ends are known as the N-terminus and

C-terminus, respectively. Proteins are synthetized in cells starting from the N-terminus by adding additional AAs to the free carboxyl group. The remaining atoms, i.e., atoms that are not involved in formation of peptide bonds, particularly on the side chains may be involved in other bonds that implement shape and function of the protein. In most living organisms there are 20 different types of AAs, each type with specific side-chain which specifies its physicochemical properties. Table 2.1 lists 20 standard AAs and some of their properties. Beside the 20 AAs, some organisms have two more AAs: Selenocysteine, and Pyrrolysine. We concentrate on proteins with the standard set of AAs, as the two non-standard AAs are found in a limited number of organisms which are not relevant to this work.

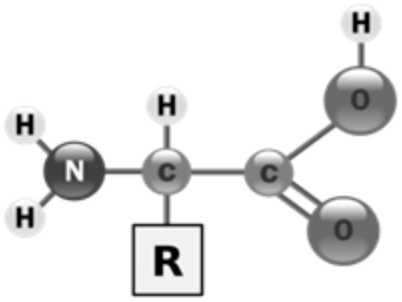


Figure 2.1: Atomic structure of an amino acid. Schematic picture of an amino acid. Balls represent atoms, connecting lines represent covalent bonds, and square labeled "R" represents a side chain. Source: Wikipedia

Table 2.1: The list of 20 standard amino acids along with their selected properties.
The table gives abbreviated names of 20 standard AAs along with chemical composition of their side chains, they selected biochemical properties, and occurrence in proteins. AA properties include annotation whether they have positive(+)/negative(-) charge, and whether they are polar(P), hydrophobic(H), aliphatic(Al), and aromatic(Ar), according to (Livingstone & Barton, 1993).

Amino Acid	Abbr.	Side chain	Characteristics				Occurrence
Alanine	Ala, A	$-CH_3$			H		7.8 %
Arginine	Arg, R	$-(CH_2)_3NH-C(NH)NH_2$	+	P			5.1 %
Asparagine	Asn, N	$-CH_2CONH_2$		P			4.3 %
Aspartate	Asp, D	$-CH_2COOH$	-	P			5.3 %
Cysteine	Cys, C	$-CH_2SH$		P	H		1.9 %
Glutamate	Glu, E	$-CH_2CH_2COOH$	-	P			6.3 %
Glutamine	Gln, Q	$-CH_2CH_2CONH_2$		P			4.2 %
Glycine	Gly, G	-H			H		7.2 %
Histidine	His, H	$-CH_2-C_3H_3N_2$	+	P	H	Ar	2.3 %
Isoleucine	Ile, I	$-CH(CH_3)CH_2CH_3$			H	Al	5.3 %
Leucine	Leu, L	$-CH_2CH(CH_3)_2$			H	Al	9.1 %
Lysine	Lys, K	$-(CH_2)_4NH_2$	+	P	H		5.9 %
Methionine	Met, M	$-CH_2CH_2SCH_3$			H		2.3 %
Phenylalanine	Phe, F	$-CH_2C_6H_5$			H	Ar	3.9 %
Proline	Pro, P	$-CH_2CH_2CH_2-$					5.2 %
Serine	Ser, S	$-CH_2OH$		P			6.8 %
Threonine	Thr, T	$-CH(OH)CH_3$		P	H		5.9 %
Tryptophan	Trp, W	$-CH_2C_8H_6N$		P	H	Ar	1.4 %
Tyrosine	Tyr, Y	$-CH_2-C_6H_4OH$		P	H	Ar	3.2 %
Valine	Val, V	$-CH(CH_3)_2$			H	Al	6.6 %

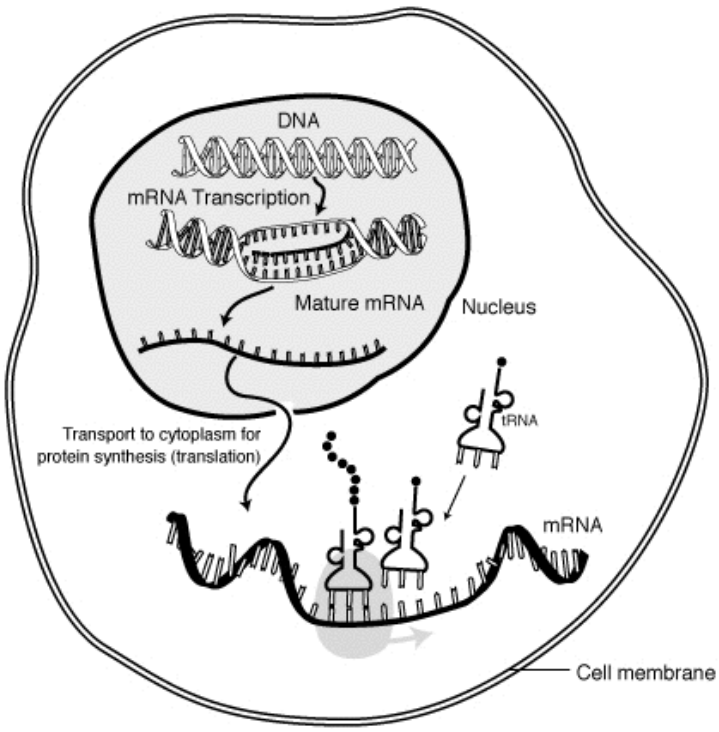


Figure 2.2: Schematic illustration of protein biosynthesis.
Source: National Institutes of Health http://en.wikipedia.org/wiki/File:MRNA-interaction.png

Biological importance of proteins has been discovered in 1926, when James B. Sumner showed that the enzyme urease was a protein (Sumner 1926). Fourteen years later, George Beadle and Edward Tatum demonstrated the existence of a precise relationship between genes and proteins (Beadle & Tatu, 1941). We now know that the information about a given protein is stored in a cell nucleus in the DNA. During protein biosynthesis a fragment of DNA, which encodes a given protein, is transcribed into messenger RNA (mRNA). The mRNA is used in next step of the biosynthesis process (in prokaryotic cells) or it first undergoes some post transcriptional modifications (in eukaryotic cells). The mRNA is next moved from the nucleus into the cytoplasm where a corresponding protein is synthesized by the ribosome through the process of translation. The mRNA is encoded using a transporting RNAs (tRNA) to produce a chain of AA, called polypeptide chain. A ribosome converts a given mRNA fragment into a corresponding protein which consists of a single polypeptide chain or a few chains, which are folded into a protein molecule that has (usually) globular shape. Figure 2.2 shows a simplified illustration of protein's biosynthesis.

After characterizing AAs composition of insulin in 1950's, Frederick Sanger started research on the field of DNA sequencing. In 1977 his research team developed so called "Sanger method", a major breakthrough which allowed long stretches of DNA to be rapidly and accurately sequenced (Sanger et al., 1977). His research earned him his second noble prize in chemistry in 1980 (shared with Walter Gilbert and Paul Berg). He is considered as the father of the field of genomics, which aims at study of genomes from different organisms. A major branch of this field is connected with sequencing genomes of different organisms and decoding AA sequences of proteins. The AA sequence is also called protein primary structure, as it represents linear order of AAs in a polypeptide chain. A recent achievement in this field was the decoding of the whole human genome, which was finished in April 2003. The ever increasing number of known protein sequences led to the birth of proteomics, which concentrates on the investigation of protein functions and structures. Proteins are responsible for almost all biological processes and they are involved in numerous function, such as transportation (transporting or storing other chemical compounds and ions), catalysis (enzymes), regulation proteins (hormones), defense (antibodies or immunoglobins), formation of structure (e.g., structure of cells), storage (they store ligands), motor functions (they convert chemical energy into movement), receptors functions (they detect signals), and signaling functions (they transmit signals), to name a few.

Proteins are often related to each other, even across different organisms, as they are governed by evolution, i.e. they may have evolved from common ancestor by different mutations introduced in DNA. To measure how closely two proteins are related we can align their sequences and look for AAs which occur at the same positions. Sometimes gaps need to be introduced in one or both sequences to allow better overall alignment, since evolution could introduce insertions or deletions. The measure of similarity between proteins is called sequence identity or sequence similarity and is represented as a count of aligned pair of AAs which are identical (sequence identity) or with similar properties (sequence similarity), divided by the sequence length. Sequences are considered as similar if they share more than 40% sequence identity. The values between 20 and 35% is so called twilight zone where the percentage of homologues sequences rapidly drops (90% for >35% and 10% for <25%) (Rost, 1999).

2.1.2 Protein Structure

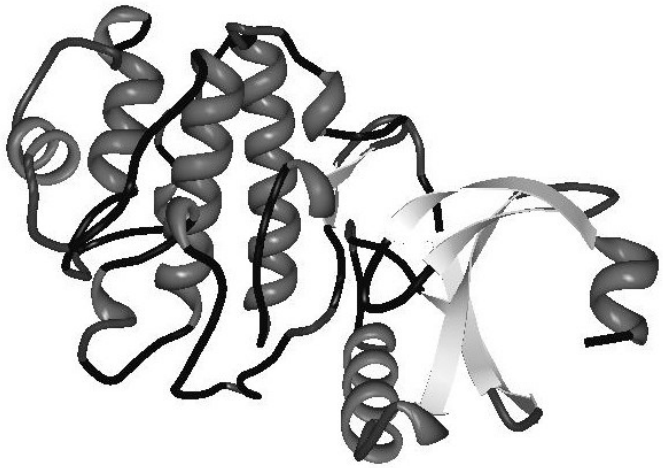
The function of proteins is usually determined by their shape, which in turn is determined by their AA sequence and environment in which the protein operates (Anfinsen, 1973). The process of formation of a specific spatial conformation/shape is called protein folding, and is guided by non-covalent interactions between protein's residues (AAs in protein sequence) and interactions between residues and solvent. A few substructures, i.e., local structural arrangements of residues, appear regularly and are called secondary structures. These structures include α helices, coils, β sheets, β bridges, turns and loops. The formation of these structures is mostly driven by hydrogen bonds between the backbone amino and carboxyl groups. A particular arrangement of the secondary structures into a compact molecule, that most often takes a globular form, is called a protein tertiary structure. The process of folding into this molecule mostly depends on hydrophobic interactions with solvent (e.g. the burial of hydrophobic residues from water) and tertiary interactions within protein, such as salt bridges, hydrogen bonds, or the tight packing of side chains. Different levels (primary, secondary and tertiary structures) of protein structure are shown in Figure 2.3.

The knowledge of the tertiary structure enables researchers to investigate biochemical and cellular functions of proteins. The protein structure is used in many areas, such as rational drug design via virtual screening (Klebe, 2000; Norin & Sundström, 2001; Guido et al., 2008; Grey & Thompson, 2010), to gain insights into various diseases (Fernàndez-Busquets et al., 2008), and to interpret interactions between proteins and other macro molecules and ligands (Luscombe et al., 2001; Ellis et al., 2007; Chen & Kurgan, 2009; Chen et al., 2011). The latter area is of particular interest as most (if not all) of protein functions are triggered or implemented through protein-ligand interactions. In most cases binding (interaction) is tight and discriminative to a narrow type of ligands, and may be examined by knowing the structure and chemical properties of a binding pocket (part of the protein structure where the interaction occurs). Interestingly, some proteins perform their function by changing their 3D conformation in response to chemical signals. This may also be found out by comparing a protein structure with and without bound ligand. Examples of such proteins include motor proteins, e.g., actin and myosin, which are involved in muscle movement.

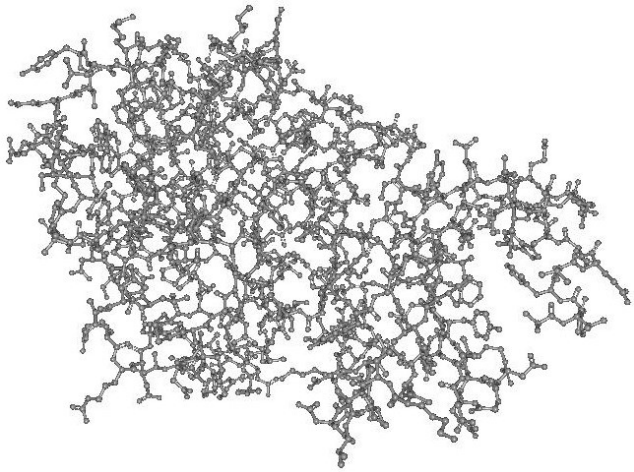
A

```
>2YER
MAVPFVEDWDLVQTLGEGAYGEVQLAVNRVTEEAVAVKIVDMKRAVDCPENIKKEICINKMLNHENVVKFYGHRREGNI
QYLFLEYCSGGELFDRIEPDIGMPEPDAQRFFHQLMAGVVYLHGIGITHRDIKPENLLLDERDNLKISDFGLATVFRYN
NRERLLNKMCGTLPYVAPELLKRREFHAEPVDVWSCGIVLTAMLAGELPWDQPSDSCQEYSDWKEKKTYLNPWKKIDSA
PLALLHKILVENPSARITIPDIKKDRWYNKPLKKGAKRP
```

B



C



D

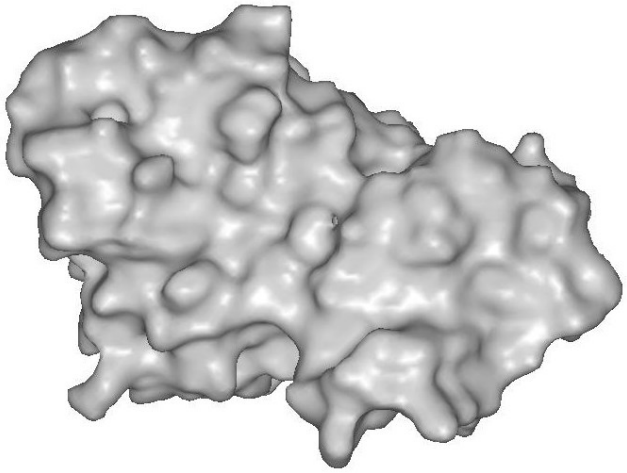


Figure 2.3: Primary, secondary and tertiary structure representations of a protein
Panel A shows primary structure (AA sequence) of the Serine/Threonine-protein kinase (PDBID: 2YER). Panel B represents 3D conformation of secondary structures which are color coded as follows: black – coils, dark grey – α-helices, and light grey – β-sheets. Panel C shows 3D structure of the protein by all atom representation. Panel D shows protein surface.

2.2 Methods for determination of proteins' 3D structure

As the knowledge of protein fold could lead to better understanding of proteins' functions and interactions, researchers explored methods to obtain their tertiary structures. The first method to succeed was X-ray crystallography, and it became the major approach for protein structure determination.

2.2.1 X-ray crystallography

In mid-19th century René Just Haüy defined crystals as regular arrays of atoms and molecules, where a single unit cell is repeated indefinitely. Since then researchers investigated possible symmetries of crystals and hypothetical atomic structures. The first actual 3D structures of crystals were derived in 1912 by Max von Laue with a help of X-ray diffraction (Nobel Prize in Physics in 1914). After that achievement the field of X-ray crystallography advanced rapidly and the structures of the crystals of smaller molecules were solved. However, proteins are hard to crystallize as they are relatively large and function in aqueous environment. The first dried protein crystal was obtained in 1926 by James Sumner (Sumner, 1926), but as it turned out the dried proteins crystals did not provide high quality X-ray diffraction patterns. The breakthrough was the first wet protein crystal in 1934 by John Desmond Bernal and Dorothy Crowfoot Hodgkin. They also had succeeded in producing an X-ray diffraction photograph of the digestive enzyme pepsin (Bernal & Crowfoot, 1934), which marks the beginning of protein crystallography. The first 3D atomic structure of a protein (Sperm whale's myoglobin) was obtained over 20 years later by Max Perutz and Sir John Cowdery Kendrew (Kendrew et al., 1958), who won the Nobel Prize in Chemistry for that in 1962.

Protein X-ray crystallography consists of three main steps: obtaining crystallized material, generating X-ray diffraction data, and analyzing the data to generate a model with 3D atomic coordinates. The first step is the most difficult, as it requires producing a pure, regular and large protein crystal, which in turns needs a milligram quantities of purified protein material (Kim et al., 2004). This challenging task is divided into following stages: production of a protein material, dissolving and purification of protein, and finally production of protein crystal. During the first stage, a protein is introduced to a crystallization vector, which is then over expressed in host cell, typically from *E. coli.*

(Structural Genomics Consortium et al., 2008). Next, the cell is destroyed and proteins are dissolved in a solvent. The solution, which contains the target protein and remaining macromolecules (including other proteins) from the host organism, is purified to contain high concentration of the selected protein. Purification is most commonly performed using various techniques which exploit differences in sizes, weight and chemical properties between protein of interest and the remaining materials. One of the procedures uses affinity tags short polypeptides attached at the end of the protein. The DNA code, which encodes the tags, is attached to the target's DNA before the target is introduced to a crystallization vector. Those tags can be used to bind proteins to immobilized protein surface or metal ions, and the remaining (non-bound) proteins are removed from the solution. The tags may later be cleaved (using protease or other enzymes which targets parts of the tags) and removed from solution as well (Waugh, 2005). The purified solution is then used to produce a protein crystal, which in turn is used in the next step where it is rotated in an intense beam of X-rays to produce diffraction patterns for each orientation of the crystal. The last step combines and analyses the generated diffraction data and computes a model of the arrangement of the atoms in the crystal. The latter two steps may be repeated several times to refine the model, and the final refined model shows a protein structure. The quality of the protein structure is expressed by many factors with crystal resolution being one of the most important and the most commonly used. Resolution is expressed in Angstroms [Å] and measures minimum distance between structural features that can be distinguished, where higher resolution, that is, smaller distance, corresponds to better structures. Hence, resolution of 1 Å is higher than resolution of 3 Å. Structures of resolution lower than 3 Å are only useful to determine general shape of the protein, whereas at resolutions higher than 1.2 Å accurate atomic coordinates can be determined (Wlodawer et al., 2008).

The process of protein crystallization is difficult because of an extremely fragile nature of protein crystals. Proteins have an irregularly shaped surface which leads to creation of large areas inside the crystal filled only with solvent, which weakens interactions between proteins in the crystal lattice. Also, the molecular variability of proteins requires a usage of unique crystallization conditions for each protein. Obtaining

diffraction-quality crystals is one of the most important bottlenecks in acquiring the structures (Chayen, 2004; Biertumpfel et al., 2005; Pusey et al., 2005; Geerlof et al., 2006). Current protocols yield crystals for approximately 30% of the input proteins (Hui & Edwards, 2003), and it is estimated that only about 2–10% of pursued protein targets yield high-resolution protein structures (Service, 2005). Furthermore, more than 60% of the cost of structure determination is consumed by the failed attempts (Slabinski et al., 2007a). At the same time, crystallization is characterized by a significant rate of attrition and is among the most complex and least understood problems in structural biology (Hui & Edwards, 2003).

2.2.2 Other experimental methods to solve protein 3D structure

The protein structures are stored in the Protein Data Bank (PDB) (Berman et al., 2000), which is the biggest, world-wide public database of 3D structures of proteins, DNAs and RNAs. PDB was founded in 1971, initially with only 7 structures, and since then the rate at which protein structures are solved continued to grow exponentially. Currently, PDB holds almost 90,000 protein structures. Except for the X-ray crystallography, there are a few other methods to solve protein structure. The most popular are Nuclear Magnetic Resonance (NMR) and Cryo-Electron Microscopy (cryo-EM). However, over the last forty years the X-ray crystallography became the main method to solve 3D protein structures; it was used to obtain 89.1% (~79,500) of the proteins deposited to PDB, whereas the second most popular method, NMR, was used to obtain "only" around 10% structures (~9,000). Together these two methods account for deciphering almost 99.1% of all protein structures deposited in PDB (as of 30th June 2013, source: http://www.rcsb.org/pdb/statistics/holdings.do). The NMR method, similarly to X-ray crystallography, also requires a sample of purified protein to work and is hard to use for bigger proteins, i.e., more than 70 kilo Daltons (kDa). Moreover, the process of obtaining 3D models by NMR is not fully automated and requires a highly trained human specialist. The third most popular method for protein structure determination, Cryo-EM, is a novel approach that utilizes electron microscopy. Because of the relatively harsh environment of the Cryo-EM experiment (high vacuum and high radiation), protein must be rapidly frozen to form a frozen-hydrated state which protects it from damage. Such prepared protein is then subjected to EM and a set of

images of rotated sample is used to create the 3D model. Although cryo-EM enables investigation of larger proteins, the resolution of these models is rather low.

2.2.3 Comparative modeling

Although PDB currently stores almost 90,000 protein structures, we know sequences for many more proteins. RefSeq (Pruitt et al., 2009) database currently includes 31+ million non-redundant (different) proteins' sequences. Moreover, the gap between the known protein sequences and solved 3D structures is widening. Since experimental methods for structure determination are expensive and their success rates are relatively low, a cost effective and fast computational (so called *in silico*) methods are being developed. A straightforward approach is to model protein folding following laws of physics, so called *de novo* protein structure prediction. However, in spite of decades of efforts to develop these methods, the structure that they generate are still not satisfactory and these methods are computationally very demanding (Moult et al., 2011). The second option to computationally predict protein structure is so called comparative protein modeling. In this approach, methods relay on previously solved structures of similar proteins, so called templates. These methods work well since in spite of the existence of a large number of sequences, there is a finite relatively low number (between a few to few dozen thousands, depending on an estimate) of protein structural motifs that these sequence fold to (Wolf et al., 2000; Vitkup et al., 2001; Koonin et al., 2002; Liu & Rost, 2002). A successful (given sufficient sequence similarity) type of comparative modeling is homology modeling, which works based on an assumption that two proteins that have similar sequence also have similar structures. Recent study shows that homology modeling methods predict the structure accurately if a sequence identity between query protein and template is at least 30%, and have a modest chance of success for sequence identity above 25% (Ginalski, 2006; Gront et al., 2012). When a protein with a sufficiently similar sequence and known structure cannot be found, a so called threading method may be applied. This method searches for distant structural similarity, i.e., proteins that do not have similar AA sequence but may still have similar structure. The template-based comparative modeling methods continue to improve (Mariani et al., 2011), however they require a database of proteins with known, experimentally solved structure and these structures would have to cover

the entire space of protein folds; otherwise it would be (it currently is) impossible to predict/compute new protein folds using this approach.

2.3 Structural genomics

Structural Genomics is another attempt to shrink the gap between the number of known AA sequences and proteins' solved structure. It aims at lowering costs and increasing speed of experimental protein structure determination by feeding hundreds of protein sequences through standardized crystallization pipelines. SG provides structures of proteins which are later used to perform template-based structure prediction, effectively enlarging our knowledge of the protein fold space and amount of available structural templates. The SG pipelines are not being optimized for a single protein but rather they are designed to produce reasonable, high-throughput results for a larger set of diverse proteins. To implement that different crystallization conditions are tested simultaneously, which helps to (automatically) find favorable conditions. This is in contrast to a traditional approach, used by structural biologists, where work with a given protein could last many years and still was not bound to result in a solved structure. SG concentrates on a protein-family-directed structure analyses, in which a group of proteins is targeted and structure(s) of representative members are determined and used to represent the entire group (Terwilliger et al., 1998). This approach also enables certain flexibility while choosing the representative proteins, as any representative of a protein family can be often selected. To this end, target selection, which is a computational process of limiting candidate proteins to those that are tractable and of unknown structure and prioritizing them according to an expected interest and accessibility (Brenner, 2000), is used.

The first efforts of SG, which were undertaken around year 2000, involved creation of the Protein Structure Initiative in the United States, a multi-center project which included four large-scale centers and six specialized centers; and similar centers in Canada, Israel, Japan, and Europe. These centers solved about the same amount of protein structures as the traditional laboratories by 2004/2005 (Chandonia & Brenner, 2006). Also, the SG's family oriented target selection addressed the drop in speed of depositions of novel structures, i.e. structures with are different to the structures already known (Levitt, 2007). SG centers also lowered the costs of structure

determination; in 2006 the cost of solving a structure at the most efficient SG center in the United States was equal to about 25% of the estimated cost when using the traditional methods (Chandonia & Brenner, 2006). Another more recent study shows that the production-line approach taken at the PSI centers reduced the average cost of solving structures from ~$250,000 apiece in 2000 to ~$66,000 in 2008 (Service, 2008).

The biggest SG project, PSI, started around 2000 and was divided into phases. So far three phases have been defined: PSI I (2000 – 2005) which concentrated on determination of feasibility of high-throughput structure determination methods; PSI II (2005 – 2010) which implemented methods developed during first phase, and most recently an ongoing third phase PSI:Biology. In the first two phases, the target selection concentrated on representatives from large, structurally uncharacterized protein domain families, and from structurally uncharacterized subfamilies in very large and diverse families with incomplete structural coverage (Dessaill et al. 2009). This approach enlarged the coverage of the protein fold space and potentially improved the computational template-based methods. However, the structure determination of sometimes biologically "not-very–relevant" targets, and targets which lack functional annotation resulted in a criticism by the biological/structural biology community (Service, 2008). To address this criticism, the new PSI:Biology phase has started in late 2010. This phase includes SG centers that will continue high-throughput structure determination and centers which specialize in structure determination of very difficult to solve and under-represented membrane proteins. Importantly, in this phase some of the targets are defined through community-based nomination process. Overall, many other targets will still follow the "classical" target selection process and would benefit from tools that support target selection. The less flexible community-based nomination process would require even more sophisticated tools which would suggest/score the crystallization protocols that are most likely to solve the nominated targets.

Over the years SG centers have contributed to development of hundreds of methods and technologies which improve crystallization protocols (Joachimiak, 2009). For example, by reducing the amount of protein needed for crystallization, which was achieved by exploiting nanovolume microfluidic environment (Maeki et al., 2012; Gerdts et al., 2006), and by enabling crystallization of proteins that can be obtained in small

amounts. Moreover, the need for the smaller amount of purified solution together with improvements in purification (Kim et al., 2011), which resulted in production of larger quantities of purified solution, enabled SG centers to screen more (500 – 1,000 screens) crystallization conditions for each target. Another example that illustrates progress in crystallization protocols includes improvements that were made to a source of X-ray radiation. Current, third-generation, synchrotron facilities not only produce radiation with excellent properties (bright, intensive and highly polarized light) but also introduce improved design of the optics, angle measuring instruments (kappa-geometry goniometer) and beamline control software. This allows for optimization of radiation properties for each crystal which enables determination of structures from lower quality and smaller crystals (Smith et al., 2012). These and other, omitted for the sake of brevity, advances continue to improve structure determination and make previously unsolvable targets solvable. This progress is ongoing, as more and more new technologies are being developed and applied; one significant example is a recent usage of X-ray laser to determine protein structure (Redecke et al., 2013). This advance was selected by Science magazine as one of one of the ten most important scientific breakthroughs in 2012, quoted to able "to decipher proteins that conventional X-ray sources cannot".

2.3.1 Crystallization data sources

Challenging aspects of crystallization as well as the increasing number of crystallized proteins led to the development of databases that record information concerning crystallization attempts. The importance of these efforts was advocated in 2000 by Raymond Stevens who said that "industrial-scale efforts will lead to the generation of knowledge bases that will be mined to expand our understanding of the techniques used in protein crystallography. These efforts will act as 'learning factories', in which successes and failures will be used to continually improve the technology for high-throughput protein crystallography" (Stevens, 2000). The development of the databases was fuelled by generation of large and well annotated experiments by SG centers, such as proteomes of *Methanobacterium thermoautothrophicum* (Christendat et al., 2000), or *Thermotoga maritime* (Lesley et al., 2002). To the best of our knowledge, the first such initiative was the PRESAGE database, which included annotations indicating current experimental status, structural predictions, and suggestions (Brenner et al., 1999). At

that time many SG consortia have established their own on-line progress reports that contained details and current experimental status of their targets. Examples include SPINE (Structural Proteomics in the NorthEast) (Bertone et al., 2001; Goh et al., 2003), Integrated Consortium Experimental Database (IceDB) (Chance et al., 2002), and Sesame (Zolnai et al., 2003); more detailed list is included in Table 1 in (Rodrigues & Hubbard, 2003). Distributed on-line progress reporting databases were gradually centralized in TargetDB (Chen et al., 2004), which was launched in July 2001 and which builds upon the work on the PRESAGE database. TargetDB was maintained by PDB and serves as a primary target registration database for structural SG project worldwide. It consolidated data from 28 SG centers in USA, Canada, Germany, Israel, Japan, France and UK, including 9 PSI centers.

However, up to 2004 the databases stored only information about successful crystallization attempts. The lack of information about failed attempts soon became a major bottleneck for studying protein properties that are associated with propensity for crystallization. This was noted in 2003 by Rodrigues and Hubbard who said “as structural genomics projects evolve, valuable experimental data will be accumulated, thus presenting researchers with a unique opportunity to establish improved predictive methods for a protein’s chemical and physical behavior based on its amino acid sequence. It is essential for laboratories producing such data to keep track of both ‘successful’ and ‘unsuccessful’ results, so that these can be fed back into the structural determination pipeline through the improvement of the target selection procedures” (Rodrigues & Hubbard, 2003). The negative/failed data is important since it allows building computational tools that can differentiate between the successful and unsuccessful attempts, and which can predict which attempts are more likely to succeed. To this end, TargetDB was extended in 2004 to create PepcDB (Protein Expression Purification and Crystallization DataBase) (Kouranov et al., 2006) that collects more detailed status information and the experimental details of each step in the protein structure production pipeline. This database stored a complete history of the experimental steps in each production trial besides describing the current target production status. PepcDB recorded status history, stop conditions, reusable text protocols and contact information collected from 15 SG centers in USA. To improve

access to experimental data and to support PSI:Biology phase, in 2012 TargetDB was merged with PepcDB and a new resource, TargetTrack, was created. TargetTrack records information from 17 centers from USA and one from South Africa and allows/encourages depositions from other centers or individual researchers. The new database improves access to experimental data, and introduces a few upgrades to better describe targets and crystallization protocols. One of particularly useful upgrades is the change of concept of protein structure target from individual protein sequences into multi component targets including assemblies of macromolecular sequences and/or ligands (see http://sbkb.org/tt/about.html for more details).

Although these databases now record the unsuccessful attempts, researchers must be careful when using these resources since the actual failure of an experiment may not be represented correctly in the database. The lack of success, i.e., lack of diffraction quality crystals, can result from the actual failure of one of the crystallization steps or for example from abandoning the work due to the changed priorities or lack of funds, which is not always properly reported. Even the actual failure of crystallization could occur due to an error or usage of wrong crystallization protocol, and not due to the actual difficulty to form a protein crystal. This inherent noise means that computational methods build using these data cannot be perfect, as this would mean that they overfit the noise.

2.3.2 Crystallization propensity studies

The high-throughput approach taken by the SG centers requires the target selection. This motivates the design of computational methods that (accurately) predict crystallization outcome. Such methods help with selecting targets which are more likely to produce diffraction quality crystals and hence reduce the crystallization costs by potentially decreasing the number of the failed attempts. Also, protein properties that are found to correlate with the crystallization outcomes could lead to the development of more accurate target selection processes and improved crystallization protocols. In early days such studies were performed by researchers at SG centers who used their own data to draw conclusions. One of the first attempts took place in 2000, when Christendat and coworkers proposed a decision tree to predict solubility from a protein sequence (Christendat et al., 2000). Later on they developed SPINE which was an integrated tracking database and a data mining method for identifying feasible targets.

Each protein deposited in this database was described using information related to the experimental progress (e.g., expression level, solubility, ability to crystallize) and 42 descriptors of the underlying protein sequence (amino acid composition, secondary structure, etc.). SG project on *Plasmodium falciparum* in 2003 found new protein characteristics important in crystallization, such as the presence of transmembrane helices, low-complexity regions, and coiled-coil regions (Rodrigues & Hubbard, 2003). In 2004, data from TargetDB were used to develop a new decision tree-based predictive model, which revealed several new protein features that influence the feasibility of target protein chain for a high-throughput structure determination (Goh et al., 2004). These features included conservation of the sequence across organisms, composition of charged residues, occurrence of hydrophobic patches in the sequence, number of binding partners, and chain length. In the same year, the study of *Thermotoga maritime* by a team from the Joint Center for Structural Genomics revealed several key features that correlate with the crystallization output, such as isoelectric point (pI), sequence length, average hydropathy, low-complexity regions, and the presence of signal peptides and trans-membrane helices (Canaves et al., 2004). The isoelectric point is calculated from the protein sequence and was also used to develop a method that suggests optimal pH ranges for crystallization screening (Kantardjieff & Rupp, 2004; Kantardjieff et al., 2004). Another study, which was conducted at the Center for Eukaryotic Structural Genomics, used disorder prediction algorithms to analyze the impact of intrinsic protein disorder on crystallization efficiency (Oldfield et al., 2005). The Berkeley Structural Genomics Center has utilized several protein features including length of the sequence and predicted transmembrane helices, coiled coils, and low-complexity regions to eliminate targets predicted to be intractable for the high-throughput structure determination (Chandonia et al., 2006).

2.3.3 Prediction of crystallization propensity

Following crystallization propensity studies, a several approaches which aimed at the prediction of crystallization propensity based on protein AA sequence were developed. These so called classification algorithms learn a prediction model utilizing a given labeled data, e.g. a set of protein labeled to reflect a given protein crystallization outcome (successful/unsuccessful), and provide prediction of label (crystallization outcome), for unseen/new samples. The input data for such algorithms must be

represented as fixed-length arrays of numbers. Hence, authors of these methods had to develop a way to transform variable-length AAs sequence into a fixed-size set of numerical features, using information which can be extracted from the sequence. A plain example of one of such features is the count of occurrences of a given AA in the sequence. Classification algorithms and the underlying machine learning concept are discussed in more detail in section 2.4.1.

The first machine learning algorithm capable of predicting crystallization output, SECRET, was developed in 2006 (Smialowski et al., 2006). This method used a relatively simple feature-based sequence representation, i.e., content (fraction in a sequence) of individual AA types, di-, and tri- peptides, using the 20 AA types and grouping AAs into sets that have similar properties. SECRET is limited to perform prediction only for proteins with length between 46 and 200 AAs. In the same year, Overton and Barton created OB-Score, a normalized scale for SG target ranking, based on only two features: pI and hydrophobicity (Overton & Barton, 2006). In 2007, two new methods were introduced, CRYSTALP (Chen et al., 2007),and XtalPred (Slabinski et al., 2007a, 2007b). CRYSTALP uses features derived from collocations of AAs pairs and has similar restriction for protein length as the SECRET method. XtalPred was created in the Joint Center for Structural Genomics to map/mimic the work performed by structural biologists. It is a white-box (human readable) model which uses nine biochemical and biophysical features of an input protein and the corresponding probability distributions, estimated based on data from the TargetDB database. The features include protein length, molecular mass, gravy and instability indices, extinction coefficient, pI, content of Cys, Met, Trp, Tyr, and Phe residues, insertions in the alignment compared to homologs in a non-redundant database of protein structures, predicted secondary structure, predicted disordered, low-complexity and coiled-coil regions; and predicted transmembrane helices and signal peptides. In this model, individual probabilities generated based on the above-mentioned features are combined into a single crystallization score, which is used to assign one of five crystallization classes: optimal, suboptimal, average, difficult, and very difficult. In the following year, Overton and colleagues upgraded their OB-score method to create ParCrys (Overton et al., 2008), a classification algorithm which uses kernel based classifier and feature vector consist of features from OB-score and

composition of selected AAs (Ser, Cys, Met, Gly, Tyr, and Phe). In 2009, two more algorithms were developed by our group, CRYSTALP2 (Kurgan et al., 2009), and MetaPPCP (Mizianty & Kurgan, 2009); they are described in Chapter 3. In the same year, PXS (Price et al., 2009) was designed at the Northeast Structural Genomics Consortium (NESG). The study that introduced this method shows that crystallization propensity depends primarily on the prevalence of well-ordered surface epitopes. More specifically, the authors show that crystallization propensity can be computed from the knowledge of predicted disordered regions, side-chain entropy of predicted exposed residues, the amount of predicted buried Gly and the fraction of Phe in the input sequence.

In the last three years, four more methods were introduced by other research groups, which include SVMCrys (Kandaswamy et al., 2010), MCSG-Z score (Babnigg & Joachimiak, 2010), XANNpred (Overton et al., 2011), and RFCrys (Jahandideh & Mahdavi, 2012). SVMCrys uses 116 features derived from predicted secondary structure and an SVM classifier. MCSG-Z score was designed on data from the Midwest Center for Structural Genomics and uses OB-score recalculated on their data and AA indices derived from AAIndex1 database (Kawashima et al., 2008). XANNpred is a pair of artificial neural networks, that take 428 features, which include mono- and di-peptide composition, pI, hydrophobicity, fraction of predicted strand and helix residues, fraction of predicted disorder, sequence length, fraction of predicted transmembrane regions, and molecular weight, as the input. The RFCRYS, depending on dataset it was applied to, uses 31 or a large set of 1341 features, which are based on AAs, di-, and tri-peptide, and pseudo AAs compositions.

Table 2.2: Overview of the existing protein crystallization propensity predictors.

Method name	Publication	Notes
SECRET	(Smialowski et al., 2006)	Protein length restriction; features based on AAs, di- and tri-peptides compositions.
OB-Score	(Overton & Barton, 2006)	Model based on only on two features: pI and hydrophobicity.
CRYSTALP	(Chen et al., 2007)	Protein length restriction; features based on AA composition and collocations of AAs pairs.
XtalPred	(Slabinski et al., 2007a)	White-box model; features based on selected AA composition, protein properties derived from the AA sequence, predicted structural characteristics, as well as homology to solved structures.
ParCrys	(Overton et al., 2008)	Upgrades OB-Score by adding features based on selected AA composition.
CRYSTALP2	(Kurgan et al., 2009)	Upgrades CRYSTALP by adding features based on selected tri-peptide collocations, hydrophobicity and pI
MetaPPCP	(Mizianty & Kurgan, 2009)	Meta predictor based on output of CRYSTALP2, XTalPred, and selected XtalPred inputs.
PXS	(Price et al., 2009)	Features based on predicted structural information, selected AAs composition, and side-chain entropy of exposed residues. Some features are based on combination of AA composition and solvent accessibility.
SVMCrys	(Kandaswamy et al., 2010)	Features based on predicted secondary structure.
MCSG-Z score	(Babnigg & Joachimiak, 2010)	Features based on recalculated OB-score and selection of AA indices.
XANNpred	(Overton et al., 2011)	Features based on mono- and di-peptide composition, pI, hydrophobicity and predicted structural information
PPCPred	(Mizianty & Kurgan, 2011)	First predictor to predict probable cause of crystallization failure; features are based on predicted protein structural information, selected AA composition, and AA indices. Some features uses information of AA compositions and values of AA indices combined with predicted solvent accessibility.
CRYSPred	(Mizianty & Kurgan, 2012)	Features based on predicted structural information, and selected AA indices. Some features uses information of AA indices combined with predicted solvent accessibility.
RFCrys	(Jahandideh & Mahdavi, 2012)	Features based on selected AAs, di-, and tri-peptide, and pseudo AAs compositions
fDETECT	N/A	Features based on selected AA composition, AA indices, instability index, and predicted sequence complexity regions.

At the same time three methods were designed by our lab, PPCPred (Mizianty & Kurgan, 2011), CRYSPred (Mizianty & Kurgan, 2012), and recently fDETECT. CRYSpred was designed to evaluate the impact of predicted protein structural characteristics, and is described in more details in section 3.3.3. PPCPred, which is described in Chapter 4, is a first-of-its-kind method that predicts not only the crystallization outcome, but it also provides details about the most probable cause of failure for the unsuccessful crystallization attempt. Finally, in response to deterioration of prediction performance of these methods over time, we recently developed a novel accurate and time-efficient method, which is named fDETECT. We utilized fDETECT to analyze crystallization propensity and the resulting structural coverage for close to two thousands proteomes. This is described in Chapter 5. All available methods are summarized in Table 2.2.

2.4 Background on computational methods

2.4.1 Machine learning

Machine learning is a branch of artificial intelligence which concerns design of algorithms that process data and act based on underlying data patterns and dependencies. One of subfields of machine learning is classification, which is a supervised learning technique. A classification algorithm learns from historical data for which labels (e.g. outcomes of experiments) are assigned to generate model that can be used to predict labels for samples which were not used to train the algorithm (e.g. outcomes of future experiments). In case of this project, each protein is a data point labeled by the outcome of crystallization experiment (e.g., crystallizable or non-crystallizable).

In classification, selection of data, which are used to train/build and evaluate the classification model, is of primary importance. After a given classification model is designed, it should be tested on independent (never used to train the model) data for which labels are known. Classifier's predictions are compared against the known labels to evaluate predictive quality (correctness) of the model. A common practice is to divide the available data instances into two subsets, where one is used for training/building of the model (training dataset) and the other for testing the predictive performance of this model (test dataset). However, the data being used not only defines how the classifier

performs but also provide an insight whether the model generalizes well enough to be applicable for the whole population or whether it is good only for a subsample of the data that is similar to the data used to train this model. This problem is known as sample selection bias and it has studied for many years. A sound solution was proposed (Heckman, 1979) and, in short, it stresses the usage of data which comes from entire population and/or to introduce corrections to predictions that would make predictions which comes from a sample applicable to the whole population. In this work we are using data available in PepcDB, which in turn includes the data from multiple SG centers. As it was mentioned before, the SG centers are not interested in individual proteins but in finding representatives in every protein family. Therefore the protein samples come from a wide range of protein families providing an unbiased coverage of the entire protein space. Another important factor when dealing with data based on protein sequences is that the proteins must be properly selected. In many cases proteins may differ between each other only by a few mutations in their sequence, and usage of such proteins in both training and test set may result in an over-estimated predictive performance; the test is performed on samples which are (too) similar to training samples. In such cases it may be enough to use alignment methods to find similar sequences. To this end, a common practice is to reduce similarity of sequences between training and test datasets. To achieve that, we introduced a threshold value for the sequence identity between proteins in our datasets, which depending on dataset is between 20 to 30%. These threshold values, as mentioned in section 2.1, are sufficient to assure sufficient sequence dissimilarity (Rost, 1999), i.e., similarity at which sequence alignment does not provide an accurate answer. A given dataset is said to include proteins below a given threshold, expressed as percentage value of sequence identity, if every pair of sequences in that dataset has sequence identity below this threshold. The final consideration in terms of data selection is the ongoing research and advances in X-ray crystallography, which could, over few years, change the result of structure determination, by solving structures of proteins which were unsolvable a few years ago. To this end, we are using data which spans only a few years back in history. That also motivates design of new algorithms every few years to maintain high predictive quality.

Another important step in designing the classification models is the selection of the classification algorithm that generates the model. Dozens (possibly hundreds) of these algorithms have been designed, and the most well-known include Naïve Bayes (John & Langley, 1995), Decision Trees (Quinlan, 1993), Logistic Regression (Schaefer et al., 1984), and Support Vector Machines (SVMs) (Vapnik, 1995). SVM was listed among the top 10 machine learning algorithms and is considered as a must-try algorithm that provides strong predictive quality (Wu et al., 2007). This algorithm is also widely used in related research (Mizianty & Kurgan, 2011; Idicula-Thomas et al., 2006; Smialowski et al., 2007; Magnan et al., 2009; Kandaswamy et al., 2010; Mizianty & Kurgan, 2012). In our work, to build our classification systems we use WEKA workbench (Hall et al., 2009), which is a collection of popular machine learning algorithms, and a popular implementation of the SVM algorithm, LIBSVM (Chang & Lin, 2011).

The existing classification algorithms require arrays of numbers as their inputs, where each data point is represented by a numerical vector of the same length. That means that protein sequences, which vary in length, must be transformed into such fixed-length array. The vector describing each data point (protein sequence in this research) is called a feature vector and each number in this vector represents a different feature/characteristic of a given protein. A simple example of features extracted from a protein sequence is the AA composition, which quantifies content/fraction of a given AA type in a protein. We explore numerous different protein characteristics to find these that are related to the crystallization outcomes and exclude these that are irrelevant. This means that we deal with feature vectors of large sizes. We search for the "best" features subset, i.e., subset that provides the most accurate prediction/classification model, utilizing feature selection methods. Given the enormous number of combinations ($2^{\text{nbr of features}}$) the exhaustive/full search for the best feature subset is impossible to perform; we simply do not have the time needed to perform such experiment. Instead, we are using heuristic search techniques, which may not find a globally best feature set, but which are capable of finding good solutions in a relatively short/manageable amount of time. In most cases we perform feature selection in two steps. First, we remove features which are poorly correlated with a given outcome, and then using a given classifier (classification algorithm) we investigate selected subsets of

features to find the best combination. To estimate the correlation of features with the crystallization outcome, we use point biserial correlation coefficient, which is defined as

$$r_{pb} = \frac{M_1 - M_0}{S_n}\sqrt{\frac{n_1 n_0}{n^2}}$$

Where:

S_n is the standard deviation of the entire set;

M_0 and M_1 is the mean value for all data points in group 0 (e.g. non-crystallizable) and group 1 (e.g. crystallizable), respectively;

n_0 and n_1 is the number of data points in group 0 and group 1, respectively;

n is the total number of all data points ($n_0 + n_1$).

The selection of feature subsets is usually implemented using a computationally-efficient best first search. The features are either added one at the time starting with empty set (forward search) or they are removed one at the time starting with the set of all features (backward search). A given feature is added/removed if this action improves predictive quality obtained by a classifier that uses the selected features when compared with the set of features before the addition/removal.

Another important factor in designing a classification method is selection of a well-performing (giving good predictive quality) set of parameter values for a given classifier; these parameters affect how the underlying model is generated. This is particularly important for the SVM method. In this work (and many other related works) the selection of parameters, a.k.a. parameterization, is performed by a grid search across parameter space. We select a combination of parameter values (from the considered grid) that results in the best predictive quality.

A typical machine learning experiment consists of the following steps:

1. **Data acquisition** – design of protocol for protein acquisition and labeling, acquire the data, and remove redundant (too similar) proteins
2. **Features generation** – generation of proteins' characteristics using AAs sequence with combination of external knowledge sources (e.g., AA indices describing AAs' hydrophobicity, predicted secondary structure, etc.)
3. **Generation of model** – Model design consists of several sub-steps. Usually several models are designed concurrently on the training dataset using different classifiers and feature selection algorithms. At the end, the final/best performing model is selected from among them. This step may be divided into following sub-steps:
    a. **Features filtration** – removal of features with low correlation with labels/outcomes
    b. **Classifier parameterization** – parameterization of selected classifiers for the full (or selected representative) set of filtered features.
    c. **Feature selection** – heuristic search for the "optimal" feature subset for a given classifier using a selected measure of predictive quality.
    d. **Classifier parameterization for feature subsets** – parameterization of each classifier and corresponding feature subset.
    e. **Evaluation of the experiment** – evaluation on the training dataset of each classifier–feature subset pair, and selection of the best model.
4. **Analysis of the selected model** – analysis of selected features and validation of the model performance on the test dataset.

Step 4 is usually performed using one or more test datasets, while step 3 is performed using a cross-validation on training dataset (see section 2.4.3). Sometimes after achieving a non-satisfactory result (e.g., predictive performance below an expected value, model that is too large/complex, etc.) a feedback loop in executed, i.e., we come back a few steps to redesign the approach in order to improve the results.

2.4.2 Evaluation of predictive performance

The evaluation of the correctness of predictions is usually performed per-protein at two levels, one for a binary outcome that evaluates the ability of a given classifier to perform decision (e.g. to crystallize or not), and the second that evaluates a numeric confidence score generated by a classifier for its binary prediction (which represents "probability" of the correctness of this prediction). The latter evaluation tells whether high/low probability values are in fact associated with the corresponding binary outcomes/labels.

The binary evaluation divides predicted samples into two categories: positives and negatives. In our case positive trials are defined as the crystallizable proteins (or a protein which passes a given step in crystallization pipeline), whereas negatives are the proteins which fail to produce diffraction quality crystals (or fail to pass a given step in crystallization pipeline). Based on the output of a binary classifier, there are four possible outcomes (two outcomes for each of two classes), including true positives (TP; correctly predicted positive trials), false positives (FP; negative trials that were predicted as positives), true negatives (TN; correctly predicted negative trials), and false negatives (FN; positive trials that were predicted as negatives). These outcomes are represented up using a confusion matrix; see Table 2.3. The quality of prediction is estimated by a variety of scores computed from this matrix. The most commonly used scores include accuracy, Matthew's correlation coefficient (MCC), sensitivity and specificity.

Table 2.3: Graphical representation of a confusion matrix.

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

Accuracy computes the ratio of correct predictions to all predictions; the higher the ratio the more accurate the predictions are. Accuracy is expressed as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

For datasets where distribution of samples between classes/outcomes is heavily unbalanced, i.e., one class label is much more abundant in the dataset than the other(s),

the accuracy may be misleading as it may be insensitive to the predictions on the smaller class(es). To this end, MCC measure, which takes into consideration distribution of the number of samples in each class, provides a better estimate of the predictive quality. MCC is defined as

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

The MCC values range between -1 and 1 and they are equal to zero when all trials are predicted as positives or negatives.

The two abovementioned quality measures evaluate predictive quality over both labels together. To this end, sensitivity and specificity are used:

$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{TN + FP}$$

Sensitivity tells how well the positive class is being predicted, whereas specificity evaluates the predictive quality for the negative class. The higher the value of each measure the more positive/negative samples are correctly predicted.

Receiver operating characteristic (ROC) curve is typically used to evaluate the numerical confidence scores. Using each unique confidence value generated by a given classifier as a threshold, all predictions with scores that are equal or greater than a given threshold are set as the predicted positives, and all other residues are set as the predicted negatives. Next, the TP-rate = TP / (TP + FN) and the FP-rate = FP / (FP + TN) are calculated and plotted; the corresponding curve is called the ROC curve. We compute the area under the ROC curve (AUC) to quantify the predictive quality. The higher the AUC value the better the scores, meaning that predictions with higher scores are more likely to be positive, while predictions with lower scores are more likely to be negative.

2.4.3 Statistical tests

We use two statistical approaches to evaluate and compare different classifiers, cross-validation and statistical test of significance, where we evaluate hypothesis that

means of two groups (predictive quality of two classifiers) are equal. The latter approach requires also a test which checks whether the analyzed data follows normal distribution.

Cross-validation is a statistical test which estimates how well results obtained on a training dataset will generalize into an independent test sets. We partition a given dataset into *x* equally-sized subsets/folds (hence the name *x*-fold cross-validation). Then, *x*–1 of these subsets are used to form a training dataset, which is utilized to design and generate a model, and the x^{th} subset constitutes the testing dataset, which is used to perform the evaluation. This is repeated *x* times, each time choosing a different subset/fold to be the test dataset. We report an average score (usually with the corresponding standard deviation) over *x* test datasets.

We use paired difference tests to compare the predictive performance of two classifiers. The outcomes of these tests may be used to check whether improvements offered by one method over another method are statistically significant (i.e., consistent). These tests need dependent observations, e.g. measurements of predictive quality for two different classifiers for the same sample. We use the Student's paired t-test if distributions of the samples are normal; or the Wilcoxon signed-rank test if they are not normal. Distribution type was verified using the Anderson-Darling test.

Student's paired t-test evaluates differences between all pairs in populations and is given by the following equation:

$$t = \frac{\bar{X}_D - \mu_0}{\sigma_D / \sqrt{n}}$$

Where:

$\bar{X}_D$ is the mean of differences between pairs;

μ_0 is a constant to which differences are compared, in this research we use 0;

σ_D is the standard deviation of differences;

n is the number of differences.

Resulting t is the test statistic which follows the Student's t distribution with n–1 degrees of freedom.

Wilcoxon signed-rank test first ranks all absolute differences between measurements and then compares the mean ranks of population. Test statistic is the absolute value of the sum of the signed ranks and is given by the following equation:

$$W = \left| \sum_{i=1}^{N_r} \left[sgn(x_{2,i} - x_{1,i}) \cdot R_i \right] \right|$$

Where:

R_i is the rank of pair (ranked by absolute difference, ties receive a rank equal to the average of the ranks they span);

$x_{1,i}$ and $x_{2,i}$ denote the measurements;

N_r is the number of pairs with non-zero difference.

For $N_r \geq 10$ z-score is calculated and if $z > z_{critical}$ then there is statistical difference. Z-score is computed using following equation:

$$z = \frac{W - 0.5}{\sigma_W}, \sigma_W = \sqrt{\frac{N_r(N_r + 1)(2N_r + 1)}{6}}$$

For $N_r < 10$ test statistic W is compared to a critical value from a reference table.

Anderson-Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution (normal distribution in this research). Here we use a variant of the test where both the mean and the variance of population are unknown and are calculated from sample of data. Test statistic A^2 is calculated form following equation:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left(\ln \Phi(Y_i) + \ln \left(1 - \Phi(Y_{n+1-1i}) \right) \right), Y_i = \frac{X_i - \bar{X}}{\sigma}$$

Where:

X_i is the i^{th} observation;

$\bar{X}$ is the sample mean;

σ is the sample standard deviation;

$\Phi(Y_i)$ is a cumulative distribution function of Y_i for normal distribution;

n is the number of observations.

Chapter 3

Prediction of crystallization propensity

3.1 Introduction and motivation

Although X-ray crystallography is used for more than 60 years to solve protein structures, the area of study of proteins' physicochemical properties that are associated with this process is relatively young. This is due to the fact that the data concerning crystallization failures have not been collected until the beginning of XXI century. Moreover, most of the costs associated with crystallization are due to the unsuccessful attempts. As it was explained in section 2.3, the biggest obstacle to obtain protein structures via X-ray crystallography is related to production of high quality protein crystals. The SG centers, which implement high-throughput protein X-ray crystallography and use standardized crystallization pipelines, have certain flexibility in target selection. To this end, accurate computational methods which help to rank proteins according to their crystallization propensity are desired; these methods would lower the costs expended on the failed crystallization attempts. This need was first addressed around seven years ago when the first machine learning-based crystallization propensity predictor has been developed. However, as prediction of protein crystallization propensity is relatively new field of study, further investigations are required. New methods explore a wider range of machine learning algorithms and find new protein's physicochemical properties which may be related to crystallization to improve predictive quality. A comprehensive evaluation and analysis of available methods is also required.

This chapter concentrates on the development of computational methods that support target selection performed at the SG centers by accurately predicting the

propensity for the proteins' crystallization outcome. To this end, we designed three crystallization propensity predictors, as well as we performed comparative analysis of predictors that were available as of end of 2009. Parts of the work presented in this chapter were published as follows:

- information about our first crystallization propensity predictor, CRYSTALP2, was published in (Kurgan et al., 2009);
- analysis and review of crystallization propensity predictors was presented in (Kurgan & Mizianty, 2009);
- design and evaluation of the remaining two predictors, MetaPPCP and CRYSpred, were published in (Mizianty & Kurgan, 2009) and (Mizianty & Kurgan, 2012), respectively.

We first briefly introduce the datasets used in this chapter, and then we describe the three predictors. Finally, we present and discuss empirical evaluation of their predictive quality and protein's characteristics used to make predictions.

3.2 Materials

We used seven datasets to design and evaluate our methods. D418 dataset introduced in (Smialowski et al., 2006) was used as the training dataset to design CRYSTALP2, however CRYSTALP2's final model was trained on a newer FEAT dataset introduced in (Overton et al., 2008). The FEAT dataset was also used to design CRYSpred. The test datasets include TEST, TEST-RL, and TEST-NEW; these were collected based on procedures described in (Overton et al., 2008) and have reduced sequence similarity with FEAT dataset that is below the 'similar structure' thresholds for the twilight-zone proteins defined in (Rost, 1999). The exact threshold value is dependent on the length of the sequence and is at about 25%. TEST-RL is a subset of TEST dataset where all longer than 250 and shorter than 46 AAs proteins were removed; this enables comparison with CRYSTALP and SECRET methods, which are constrained to this protein chain length range. FEAT, TEST and TEST-RL datasets contain proteins which were deposited to PepcDB and TargetDB before April 2007. To compare predictors on newer depositions, we created TEST-NEW dataset which contains proteins with never

deposition dates i.e. depositions between July 2006 and January 2009. The TEST_NEW dataset was introduced in (Kurgan et al., 2009) and proteins were extracted using the same protocol as the one used to extract the FEAT and TEST datasets (Overton et al., 2008). This dataset was further randomly split in (Mizianty & Kurgan, 2009) to form TR-1500 and TEST-500 datasets which were used to build and evaluate MetaPPCP. Details about these datasets are given in Table 3.1.

Table 3.1: Summary of datasets used to develop and evaluate CRYSTALP2, MetaPPCP, and CRYSpred methods.
NC stands for non-crystallizable, C stand for crystallizable.

Dataset name	Number of proteins		Introduced in	Notes
	NC	C		
D418	192	226	(Smialowski et al., 2006)	
FEAT	728	728	(Overton et al., 2008)	
TEST	72	72	(Overton et al., 2008)	Low sequence similarity to the FEAT dataset
TEST-RL	43	43	(Overton et al., 2008)	Subset of TEST, only proteins of length between 46 to 200 AAs
TEST-NEW	1,000	1,000	(Kurgan et al., 2009)	Contains depositions newer than those in FEAT and TEST
TR-1500	750	750	(Mizianty & Kurgan, 2009)	Subset of TEST-NEW
TEST-500	250	250	(Mizianty & Kurgan, 2009)	Subset of TEST-NEW

3.3 Proposed approaches

3.3.1 CRYSTALP2

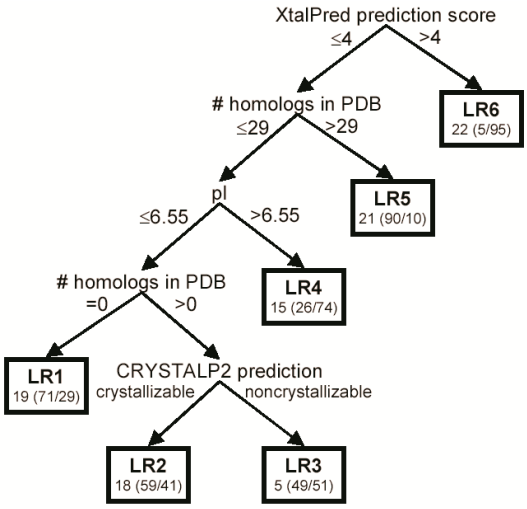
Our first method, CRYSTALP2, builds upon design of its predecessor CRYSTALP (Chen et al., 2007). 34,022 features were considered during the design, which include 2,020 features considered in CRYSTALP and 32,002 new features. Features previously used to develop CRYSTALP include AA composition (20 features), and dipeptide AA collocations, i.e., any specific two AAs which are next to each other or are separated by up to 4 different residues $AA_i\{-\}_{0\text{-}4}AA_j$ (2000 features). The novel features included tripeptide AAs collocations, where each of considered AAs may be separated by up to 1 residue $AA_i\{-\}_{0\text{-}1}AA_j\{-\}_{0\text{-}1}AA_k$ (32,000 features), isoelectric point, and hydrophobicity. A (relevant) subset of these features was selected using two rounds of correlation-based feature

subset selection method (CFSS) (Hall, 1999) using ten-fold cross-validation on the D418 dataset. CFSS evaluates the value of a subset of features by considering the individual predictive capability of each feature as well as their redundancy, selecting features which are highly correlated with outcomes and which share low correlation with each other. As a result, we selected 88 features. The features were then fed into the normalized Gaussian radial basis function (RBF) network, which is a neural network with a hidden layer based on the non-linear Gaussian kernel function (Bugmann, 1998). We utilized the RBF network implementation in WEKA.

3.3.2 MetaPPCP

Following the design of CRYSTALP2, we investigated complementarity of the four best performing crystallization predictors at that time: CRYSTALP2, OB-score, ParCrys, and XtalPred (Kurgan & Mizianty, 2009). To assess whether a meta approach, which combines results from multiple "base" predictors, would provide improved predictive quality when compared with these individual predictors, we considered an oracle predictor where we assumed that prediction is correct when at least one of the four methods is correct. The oracle was able to correctly predict around 90% of targets on the TEST-NEW dataset, compared to substantially lower success rate for individual "base" methods, which is at around 70% accuracy. This suggested that the base methods are complementary and could be combined together to improve the predictive quality. However, a simple meta approach, which was implemented using majority vote (with CRYSTALP2 as a tie-breaker), achieved relatively small improvement (~73% accuracy) over the individual base methods. Thus, we designed a more complex meta approach called MetaPPCP. During design of this meta classifier we considered features which included outputs generated by OB-Score, ParCrys, XtalPred, and CRYSTALP2 predictors, such as their predicted outcome (crystallizable vs. non-crystallizable) and the predicted confidence score (probability). We also added the information generated by the XtalPred server, which includes chain length, pI, Gravy and instability index values, average number of insertions in the alignment compared to homologs (structures with similar sequence) in the non-redundant (NR) protein database filtered at 60% protein identity, number of homologs in NR and PDB databases, and predicted percentage of coils, coiled coils, the length of the longest disorder region, transmembrane helices, and signal peptides (Slabinski et al., 2007b). During the design process, we evaluated a wide

range of classification methods implemented in WEKA platform, such as linear logistic regression, SVM, probabilistic Naïve Bayes, C4.5 decision tree, and logistic model tree (LMT) (Landwehr et al., 2005). Each of these classification models was parameterized using the full set of features and five-fold cross-validation on the TR-1500 dataset. The parameterized classifiers were used to perform two best-first search based feature selections, forward and backward. Finally, each classifier was parameterized again using the selected feature set and five-fold cross-validation on the TR-1500 dataset. As a result, we considered fifteen designs where five types of prediction models are executed on three different feature sets. Although the highest MCC (which we used to assess predictive quality) was achieved by the SVM-based design, we decided to use Logistic model tree with forward best search feature selection as it also achieved relatively high MCC value on the training dataset and used a smaller number of features. The resulting MetaPPCP method uses CRYSTALP2 and XTalPred outputs, as well as gravy hydrophaty index, pI, and number of homologs in PDB as its input features. ParCrys and OB-Score heavily rely on hydrophaty index and pI, which may explain why they were not used in our meta approach. The MetaPPCP model is presented in Figure 3.1.



Linear regression coefficients for crystallizable class[1]

Input variables	LR1	LR2	LR3	LR4	LR5	LR6
Gravy index	-1.52	-0.24	-1.16	-2.29	0.09	0.03
pI	-0.73	0.06	-0.02	-0.13	-0.2	-0.12
# homologs in PDB	-0.01	0.01	0.03	-0.02	0.00	0.00
CRYSTALP2 prediction	-0.15	-0.21	-0.21	-0.40	-0.32	-0.93
XtalPred predict. score	-0.40	0.09	-0.24	-0.30	-0.07	-0.27
error term (constant)	5.21	-0.44	0.37	1.03	2.40	1.26

[1]the coefficients for the noncrystallizable class have the same value and opposite sign

Figure 3.1: Overview of the MetaPPCP prediction model.
Figure shows the proposed model, the decision tree is shown on the left and the linear regression (LR) models from the leaf nodes are shown on the right. The left most "LR1 19 (71/29)" leaf node denotes that its corresponding linear regression is LR1, and that this node concerns 19% of the input proteins among which 71% are crystallizable and 29% are non-crystallizable.

3.3.3 CRYSpred

Following the design of the meta approach, in 2010 we worked on the CRYSpred method. The main purpose of this study was to investigate which AA indices (which quantify a wide range of physiochemical and structural properties of different AA types) are useful to predict crystallization propensity, as well as to assess whether additional information predicted from the sequence, such as solvent accessibility and disorder information, could improve classification accuracy. We explored a wide range of AA indices from the AAIndex1 database (Kawashima et al., 2008) and combined them with the solvent accessibility predicted by SPINE (Dor & Zhou, 2007). We defined a given residue as solvent exposed if its predicted relative solvent accessibility is greater than 0.25; otherwise we assume that the residue is buried. We investigate total of 531 AA indices; we excluded the indices that have missing values for any of the 20 AAs. For each index we computed following three values (which results in 1593 features):

- {AAIndex} – sum of the index values for each residue divided by the sequence length (531 features)
- {AAIndex}_exp – sum of the index values for residues predicted as solvent exposed divided by the number of the exposed residues (531 features)
- {AAIndex}_bur – sum of the index values for residues predicted as buried divided by the number of the buried residues (531 features)

We also include 7 features which are based on the disorder predicted with DISOPRED2 (Ward et al., 2004):

- DIS_RES – number of the predicted disordered residues divided by the sequence length (1 feature)
- DIS_MAX{_norm} – the length of the longest predicted disordered region (either normalized with respect to the sequence length or not) (2 features)
- DIS_AVG{_norm} – the average length of the predicted disordered regions (either normalized with respect to the sequence length or not) (2 features)
- DIS_REAL – sum of predicted disordered scores for each residue, divided by the sequence length (1 feature)
- DIS_SEG – number of predicted disordered regions (1 feature).

We applied features selection to remove features that are irrelevant or weakly-relevant to the prediction of the crystallization propensity. To filter the initial set of 1,600 features we ranked each feature according to its MCC value, based on the output

of Flexible Naïve Bayes classifier, with only one feature at the time, on five-fold cross-validation on the FEAT dataset. We selected this classifier since it allows for quick computation of the MCC value, while we had to compute 1,600*5 = 8,000 models. We use the MCC value for the isoelectric point (pI), which equals 0.286, as a cut-off threshold, i.e., the features with the MCC values < 0.286 were filtered out. The pI was selected as it is one of the features that are known to affect the crystallization (Canaves et al., 2004). Consequently, we selected 161 features that have the MCC values between 0.286 and 0.415.

We considered the SVM classifier with three popular kernel functions including Radial Basis Function (RBF), Polynomial kernel (POLY), and normalized polynomial kernel (NPOLY). We performed parameterization and feature selection based on the five-fold cross-validation on the FEAT dataset, aiming at maximizing the MCC values. The best-performing model was SVM with normalized polynomial kernel with 15 selected features, however all six designs scored similarly well with MCCs between 0.54 up to 0.57.

Among the fifteen selected features we observed a strong presence of information derived from the charge-based AA indices, which agrees with previous observation made in (Goh et al., 2003), and from the hydrophobicity-based AA indices, which concurs with the observations in several related studies(Overton & Barton, 2006; Babnigg & Joachimiak, 2010; Goh et al., 2003; Chen et al., 2007; Kurgan et al., 2009; Overton et al., 2008). The selected feature set uses two AA indices that describe AA composition, which was also used in several prior methods that predict crystallization propensity (Slabinski et al., 2007a; Price et al., 2009; Kurgan et al., 2009; Overton et al., 2008). Three AA indices described secondary structure propensities of AAs, more specifically propensity for the alpha helix conformation, which could be associated with the fact that membrane spanning regions in protein structure are often implemented with alpha helices; this information was previously found useful for the crystallization prediction (Rodrigues & Hubbard, 2003; Canaves et al., 2004; Chandonia et al., 2006). Our method also utilizes features derived from the predicted disorder, which agrees with the findings in (Slabinski et al., 2007a; Oldfield et al., 2005), and information concerning the predicted solvent accessibility, which was shown to be important in

(Derewenda, 2004; Goldschmidt et al., 2007). Overall, we observe that the selected features are well aligned with existing observations. Our predictor is arguably the first to combine these features/observations together.

Table 3.2: Summary of selected features for the CRYSpred method.
Summary of the selected 15 features used in the CRYSpred method along with their MCC values obtained by the Flexible Naïve Bayes classifier and biserial correlation on the FEAT dataset (see the "Features" section for more details). The features are sorted in the descending order according to the MCC values. The "Feature name" uses identifiers of the corresponding AA indices from the AAindex database.

MCC	Biserial	Feature name	Description	Feature type
0.397	0.349	NAKH900113	Average value of AA index describing AA composition	AA composition
0.368	0.386	KUMS000103	Average value of AA index describing distribution of AAs in the alpha-helices in thermophilic proteins	Secondary structure
0.367	0.390	KUMS000104	Average value of AA index describing distribution of AAs in the alpha-helices in mesophilic proteins	Secondary structure
0.360	-0.417	GRAR740101	Average value of AA index describing AA composition	AA composition
0.347	-0.336	DIS_MAX_norm	The length of the longest predicted disordered region divided by the sequence length	Disorder
0.343	0.361	QIAN880103_exp	Average value of AA index describing weights for alpha-helices	Secondary structure
0.325	-0.232	PARJ860101	Average value of AA index describing HPLC parameter	Hydrophobicity
0.315	0.129	WERD780101	Average value of AA index describing propensity of AAs to be buried	Solvent accessibility
0.312	-0.272	DIS_REAL	Sum of predicted disorder scores for each residue divided by the sequence length	Disorder
0.309	0.183	BIOV880101	Average value of AA index describing solvent accessibility of AAs	Solvent accessibility
0.307	0.116	BAEK050101	Average value of linker index	Disorder
0.307	0.289	COWR900101	Average value of hydrophobicity index	Hydrophobicity
0.299	-0.322	CHAM830108	Average value of AA index describing a parameter of charge transfer donor capability	Charge
0.299	0.287	FAUJ880112_bur	Average value of AA index describing negative charge for buried AAs	Charge
0.292	0.205	FAUJ880112	Average value of AA index describing negative charge	Charge

3.4 Empirical evaluation

Table 3.3 summarizes experimental evaluation performed on the TEST, TEST-RL and TEST-NEW datasets for our three predictors as well as a representative set of existing methods developed by other groups. We also provide evaluation on the TEST-500 dataset, with the exception of CRYSpred that was not evaluated on this dataset.

At the time of publication **CRYSTALP2** achieved second-best predictive quality on the TEST dataset; however, the best at that time XtalPred predictor used PDB dataset to generate its features, which overlapped with the TEST dataset and gave it a potential advantage. Comparison on the newer TEST-NEW dataset showed that CRYSTALP2 and other related methods (OB-score, ParCrys and XtalPred), are characterized by relatively similar prediction quality with MCC and accuracy values ranging between 0.39 and 0.43 and between 69.3 and 70.6%, respectively. To compare, the accuracy of a random predictor is at 50%. However, we later found that these predictors are complementary, which was exploited to propose MetaPPCP (Kurgan & Mizianty, 2009).

MetaPPCP and **CRYSpred**, outperform the other methods, including CRYSTALP2. These two methods were only directly compared on the TEST dataset, as MetaPPCP was trained on part of the TEST-NEW dataset and could be evaluated only on its remaining part, TEST-500, while CRYSpred was evaluated on the entire TEST-NEW dataset. The direct comparison on the TEST dataset reveals that MetaPPCP achieves results which are slightly better by 0.6% accuracy and 0.01 MCC. On the TEST-500 dataset, MetaPPCP outperforms the other methods by almost 8% accuracy and 0.14 MCC, whereas for the TEST dataset the improvement is smaller (1.3% and 4.8% accuracy, and 0.03 and 0.09 MCC over 3rd best XtalPred and 4th best CRYSTALP2 respectively). Our second predictor, CRYSpred, obtains the highest predictive quality on the TEST-NEW dataset, where it achieved 3% and 0.04 improvements in accuracy and MCC, respectively. We note that the accuracies on the TEST-NEW dataset are lower by about 6% for the CRYSpred and 9% for the second-best ParCrys, when compared with the results on the TEST and TEST-RL datasets. This drop in the quality of the predictions could be explained by the fact that the TEST-NEW dataset includes newer data. We also notice that all considered methods (except CRYSpred and XtalPred on TEST-RL) have higher values of sensitivity

than specificity. That means that all classifiers are less likely to miss crystallizable targets.

Table 3.3: Comparison of predictive quality between existing crystallization propensity predictors on the TEST, TEST-RL and TEST-NEW datasets.
Compared methods include: CRYSpred, MetaPPCP, ParCrys, OB-Score, XtalPred, CRYSTALP, CRYSTALP2, SECRET, and SVMCrys. The CRYSpred was trained on the FEAT dataset, MetaPPCP was trained on part of TEST-NEW dataset, whereas the results for the ParCrys, OB-Score, XtalPred, CRYSTALP, CRYSTALP2 and SECRET were obtained from the web servers and for the SVM-CRYS using the author-provided standalone application. The SECRET and CRYSTALP could be tested only on the TEST-RL dataset since they predict only sequence with 46 to 200 residues. MetaPPCP could not be compared on TEST-NEW dataset as part of it was used in its design. The results on each dataset are sorted in the descending order according to the MCC values and the best results for each quality index and each dataset are shown in bold font.

Dataset	Method	Year published	Accuracy	MCC	Sensitivity	Specificity	AUC
TEST	MetaPPCP	2009	**80.4**	**0.61**	81.7	**79.2**	0.84
	CRYSpred	2012	79.9	0.60	**81.9**	77.8	**0.85**
	XtalPred	2007	79.2	0.58	79.2	**79.2**	0.83
	CRYSTALP2	2009	75.7	0.52	79.2	72.2	0.79
	ParCrys	2008	71.5	0.45	79.2	58.3	0.75
	OB-Score	2006	64.6	0.32	87.5	47.2	0.68
TEST-RL	CRYSpred	2012	**80.2**	**0.60**	**76.7**	**83.7**	**0.86**
	ParCrys	2008	79.1	0.58	N/A[1]	N/A[1]	0.84
	XtalPred	2007	76.7	0.54	74.4	79.1	0.82
	CRYSTALP2	2009	69.8	0.40	74.4	65.1	0.72
	OB-Score	2006	69.8	0.40	N/A[1]	N/A[1]	0.71
	SECRET	2006	58.1	0.16	N/A[1]	N/A[1]	0.58
	CRYSTALP	2007	46.5	-0.07	N/A[1]	N/A[1]	N/A[2]
TEST-NEW	CRYSpred	2012	**73.4**	**0.47**	80.6	**66.5**	**0.78**
	ParCrys	2008	70.6	0.43	**83.2**	58.1	0.75
	SVMCrys	2010	70.4	0.42	78.6	62.3	N/A[2]
	OB-Score	2006	69.8	0.42	85.6	54.1	0.74
	XtalPred	2007	70.0	0.40	75.7	64.4	0.76
	CRYSTALP2	2009	69.3	0.39	76.1	62.6	0.74
TEST-500	MetaPPCP	2009	**81.0**	**0.63**	**88.9**	**73.3**	**0.88**
	OB-Score	2006	73.0	0.49	88.9	57.8	0.76
	ParCrys	2008	73.4	0.48	84.4	62.9	0.77
	XtalPred	2007	72.4	0.45	77.0	68.0	0.77
	SVMCys	2010	70.0	0.41	78.3	62.1	N/A[2]
	CRYSTALP2	2009	68.4	0.37	73.4	63.7	0.75

[1] – Results were not published and cannot be recomputed

[2] – The SVM-CRYS and CRYSTALP do not produce crystallization propensity scores (they only generate binary predictions) and thus we could not compute their AUC values.

CRYSpred achieves also the highest AUC values at about 0.85 for the older two datasets (TEST and TEST-RL), and 0.78 for the newer and larger TEST-NEW dataset. The AUC scores and the corresponding ROC curves show that CRYSpred works on average better than the ParCrys and XtalPred methods, and that these improvements hold for the majority of the range of the TP- and FP-rates, see Figure 3.2.

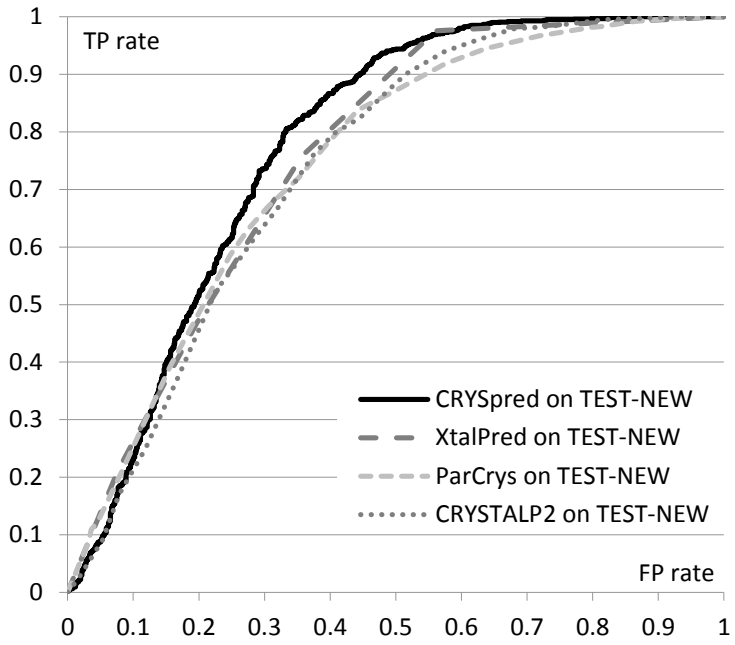


Figure 3.2: The ROC curves for the ParCrys, XtalPred, CRYSTALP2, and CRYSpred methods computed on the TEST-NEW dataset.

We also investigated the complementarity of the methods by combining multiple methods using OR operator, i.e., a given prediction is assumed correct if at least one of the methods in an ensemble provides a correct prediction. This approach allows quantifying the amount of overlap in predictions and it also estimates the upper limit of predictive quality for a potential meta-predictor that combines predictions from the individual methods. Figure 3.3 shows summary of results, in terms of achieved TPR, TNR and MCC values for all combinations of two, three, and four predictors as well as for the individual methods. We observe that certain ensembles obtain higher quality of predictions indicating a stronger complementarity. In particular combining either OB-Score and XtalPred or CRYSTALP2 with XtalPred gives better results than any other combination of two methods. Among the ensembles of three methods, the combination of XtalPred and CRYSTALP2 with either ParCrys or OB-Score works best. This observation

and the fact that OB-Score and ParCrys are the least complimentary among all pairs of predictors indicate that these two methods provide relatively overlapping outputs. Moreover, an ensemble of all four methods obtains MCC of 0.82 which is not much higher than 0.80 achieved with just three methods, showing that addition of the fourth predictor brings relatively minor improvements. Finally, we again observe that results indicate that both individual and ensemble-based predictions are characterized by higher quality for crystallizable rather than non-crystallizable proteins.

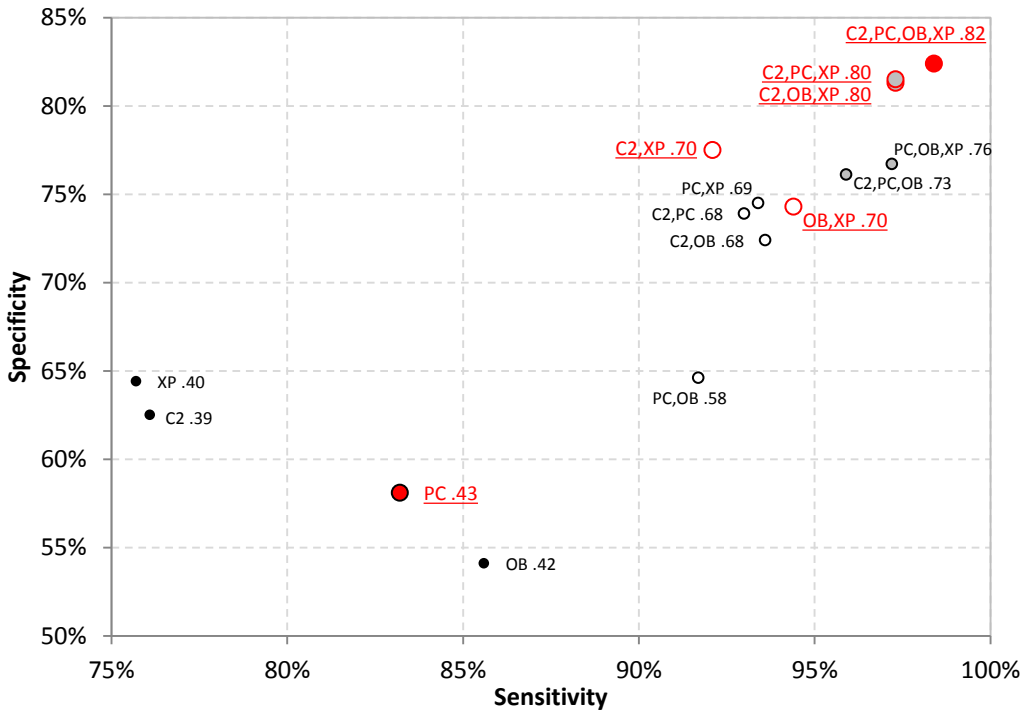


Figure 3.3: Analysis of the complementarity of crystallization propensity predictions including OB-Score, ParCrys , XtalPred, and CRYSTALP2.

Analysis of the complementarity of predictions for OB-Score (OB), ParCrys (PC), XtalPred (XP), and CRYSTALP2 (C2) methods on the TEST-NEW dataset. Each combination of 1, 2, 3, and 4 methods was applied using OR operator, i.e., a given prediction was assumed correct if at least one of the predictors predicted it correctly. Points representing combination of 1, 2, 3, and 4 methods are black, hollow, grey, and red, respectively. The x-axis/y-axis shows sensitivity/specificity values (sensitivity values are scaled between 75 and 100% while specificity values are scaled between 50 and 85%), and the labels next to markers denote a particular combination of applied predictions together with the MCC value (e.g., "PC,XP,C2 .80" means that combination of ParCrys, XtalPred and CRYSTALP2 obtained MCC of 0.8). Markers and labels in red denote the best results for a given number of applied methods.

3.5 Conclusions

We designed three accurate (at the time of publication) crystallization propensity predictors and performed empirical evaluation of their predictive performance. Our two predictors, MetaPPCP and CRYSpred, achieved the top accuracy and MCC across all datasets outperforming remaining solutions. Evaluation of the top performing methods in 2009 shows that their predictions are complementary to each other, which motivates further research as this may indicate that further improvement in predictive performance is possible.

The physicochemical characteristics (features) used by our predictors are well-grounded in the prior studies that investigated factors related to the propensity for protein crystallization, which lowers novelty of our work. However, the combined usage of those characteristics (e.g., combining certain characteristics with specific levels of solvent accessibility), or computation of certain values from predictions (e.g. normalized length of the longest disordered segment) have not been done before and we empirically demonstrated that these novel contributions are helpful in generating more accurate predictions. This motivated us to further explore possible feature space, especially using information contained in AAIndex database, and design novel features which would be based on properties which are known to be related to crystallization propensity.

Finally, our study showed that the predictive quality obtained by various methods on the newer test data is lower. We hypothesize that this could be a consequence of advances in the crystallization protocols (Fogg & Wilkinson, 2008; Grey & Thompson, 2010), which would potentially enable crystallization of previously non-crystallizable proteins. This in turn would confuse the results generated by the prediction models that were established using older data, i.e., the FEAT dataset that was proposed in (Overton et al., 2008).

Chapter 4

Prediction of outcomes from X-ray crystallography pipelines

4.1 Introduction and motivation

Based on the results from chapter 3 we drew two conclusions/hypotheses: (I) predictors trained on older datasets tend to generate worse results on newer data, and (II) there is room for further improvement in predictive quality which could be achieved by using newly designed features that are better associated/correlated with the crystallization propensity. We also realized that prediction of the crystallization propensity, however very helpful for target selection performed by the SG centers, provides limited help to crystallographers who need to crystallize a given protein, as it does not indicate the cause of crystallization failure.

To address the issue of improvements in the crystallization protocols (which render older methods to become less accurate) and inability of the existing methods to indicate a reason for the predicted failure of crystallization, we decided to design a new predictor. This method was built on newer data extracted with re-designed and improved annotation protocol. We also decided to extended our predictor to indicate the most probable obstacle in the crystallization process in case when the negative outcome is predicted (protein cannot be crystallized), besides predicting the crystallization propensity. This extended prediction provides more insights into crystallization properties of the targets, and may help researchers to modify a given target to increase chances to pass the step that leads to the predicted failure, e.g., one may introduce tags at the sequence termini to ease purification when the purification failure is predicted. This method may also be useful for researchers which are not interested in crystallization, but rather in obtaining purified protein material. Finally, we

decided to develop a novel set of features with improved correlation with crystallization propensity of proteins.

In collaboration with Drs Helen Berman and John Westbrook from Rutgers University, who are the curators of the PepcDB and TargetDB databases, we defined three steps which map the most important subsequent steps of the crystallization procedure:

i. **production of the protein material** – includes protein cloning and expression, and partially protein solubilization
ii. **purification** – which includes both protein solubilization and purification
iii. **production of diffraction quality crystals** – which includes obtaining high quality X-ray diffraction patterns

The data for the design and comparative evaluation of our method was obtained from PepcDB using a more precise and comprehensive annotation protocol, when compared with the prior works; this effort is described next. Following that, we describe the design of new method which predicts outcomes of each of abovementioned three steps, and we perform empirical evaluation of its predictive quality. Finally, we investigate and describe protein characteristics which were used to perform predictions.

The contents of this chapter were published in (Mizianty & Kurgan, 2011).

4.2 Materials

4.2.1 Annotation and datasets extraction protocol

We used PepcDB (Kouranov et al., 2006) downloaded on Nov 17th 2010, which consisted of 261,572 targets to create suitable training and test datasets. In this database target is defined either as a single protein or a collection of protein chains. Each target may have one or more associated trials, which represents a set of procedures used to crystallize a target. There are 817,099 trials in our PepcDB dataset. Each trial has information about it current status and, in case if work was finished, the stop status, see Table 4.1. The stop statuses indicate the step at which the work on a given trial was stopped and the reason of the failure or the fact that the trial produced a proper outcome. The majority of trials have the stop status field empty, which makes it

impossible to deduct the final outcome of the trial; we cannot be sure whether the experiment was finished, abandon, or is still in progress.

Table 4.1: List of stop statuses and current statuses in PepcDB.
The statuses are sorted top-down from steps earlier to further in the crystallization procedure. The current status sometimes indicates the current state of an experiment, rather than the completed activity, e.g. for the "cloning failed" stop status, the current status "cloned" does not mean that cloning was successful, but if the current status is "expressed" then cloning can be assumed successful. We disregarded "other" , "poor NMR", "mass spec failed" and "duplicate target found" stop statuses, and "other", "test target", "work stopped", "selected", "mass spec verified", "NMR assigned", "HSQC", "NMR structure" current statuses.

Class deduced from PepcDB annotation	Stop status	Current status
Production of protein material failed	sequencing failed , cloning failed	cloned
	expression failed	expressed
Purification failed	purification failed	soluble
		purified
Crystallization failed	crystallization failed	crystallized
	poor diffraction	diffraction-quality crystals
		diffraction (native diffraction-data or phasing diffraction-data)
Crystallizable	structure successful, TargetDB duplicate target found, PDB duplicate found	crystal structure
		in PDB

We extracted all trials from PepcDB with completed stop statuses listed in Table 4.1, and with the current status "in PDB" or "crystal structure", as they clearly indicate the successful crystallization attempts (step 1 in Table 4.2). Since each trial may concern more than one sequence, we considered each sequence from each trial as a separate trial. As information in PepcDB could be inconsistent we use the following 3 filtering steps to improve reliability of the target annotations.

First (step 2 in Table 4.2), we filtered the sets to remove the trials with duplicate sequences based on their stop status. For all pairs of trials with the same sequence, we removed the trial with an earlier stop status (see Table 4.1), as the second approach was successful at passing this (earlier) step. In case of two trials with the same stop status, we removed the older trial. We also filtered all non-crystallizable chains against the PDB, i.e. we remove a given chain from the set in case if this sequence occurs in the PDB (step 3 in Table 4.2). These conflicting stop statuses for the same sequences may be the results of experiments performed by different groups and or using different

crystallization protocols, where one group/protocol was able to pass the step which caused failure for another group.

In the next step (step 4 in Table 4.2), we further filtered the non crystallizable chains against all trials in PepcDB based on their current status field. We removed each non-crystallizable trial for which there is a trial with the same sequence and the current status further along the crystallization process (see Table 4.1). In this case, the current status indicates that the trial succeeded with the step (stop status). We did not include the sequences with current status, as it was mentioned before we cannot be sure of the final outcome of the experiment.

Finally (step 5 in Table 4.2), we removed all non-crystallizable trials from before Jan 1st, 2006 and after Dec 31st, 2009. We removed the older samples to accommodate for the latest advances in the crystallization protocols. For example, our analysis of the PepcDB shows that before 2006, i.e., in the first PSI phase, a large number of failures corresponded to problems with cloning, whereas after 2005 the problems with cloning subsided. The samples from 2010 could not be used since at the time of publication some of them may not be yet completed or updated in the database.

We grouped the trials which remained after this filtration into the following four classes/outcomes: production of the protein Material Failed (MF), Purification Failed (PF), Crystallization Failed (CF), and CRYStallizable (CRYS). Grouping was made based on their stop statuses; see step 6 in Table 4.2 for details. Finally, using BLASTCLUST program (http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html) we reduced the sequence identity among chains within the same class, i.e. for each class we kept only the sequences below 25% sequence identity threshold. This is consistent with the threshold used in the prior studies (Overton & Barton, 2006), but we did not reduce the sequence identity between trials from different classes.

We created 3 datasets which were used to design predictors for each of the non-crystallizable classes/outcomes (MF, PF, and CF). Each of these datasets includes trials which failed to proceed through a given step (as the samples in the negative set), and trails which passed this step (as the positive set). For the PF and CF datasets we did not include trials from the MF, and the MF and PF classes, respectively, since we do not

known whether these trials would pass the purification or crystallization steps since they did not pass the previous steps, e.g., we do not know whether the trials which failed to produce material would purify if they pass the production of the protein material step.

Table 4.2: Distribution of samples in the datasets that was used to develop and evaluate method for prediction of outcomes from crystallization pipelines.
Number of samples in each step used to create datasets (top of the table) and the sizes of the final datasets (bottom of the table where shading denotes the data aggregated for a given class label). The steps include: 1. Selecting proteins with the completed stop status; 2. Filtering out trials with the same sequence; 3. Filtering out the non-crystallizable proteins against PDB and CDB; 4. Filtering out the non-crystallizable proteins against trials in PepcDB based on their current status field; 5. Selecting trials between 2006 and 2009; 6 Assigning class labels; 7. Removing sequence identity within each class.

Step	Non-crystallizable, with the following failed stop status						Crystallizable
	sequencing	cloning	expression	purification	crystallization	diffraction	
1	508	6,222	11,223	16,457	5,123	6,391	15,412
2	244	3,490	7,252	7,819	4,093	1,283	7,006
3	243	3,470	7,225	7,641	4,087	1,267	7,006
4	240	3,216	7,152	7,462	4,087	1,267	7,006
5	27	764	3,902	4,737	3,135	1,205	4,779
6		4,693		4,737	4,340		4,779
7		2,486		1,431	849		2,408

Datasets	Production of the protein material failed	Purification failed	Crystallization failed	Crystallizable
DB_4CL	2,486	1,431	849	2,408
DB_MF	2,486		4,688	
DB_PF		1,431	3,257	
DB_CF			849	2,408
DB_CRYS	4,766			2,408

We also created the DB_CRYS dataset with the class labels/outcomes similar with the previous research in this field, which indicate the success of the entire process, i.e., production of the diffraction-quality crystals, and the dataset with 4-class annotations, which includes all 4 outcomes (DB_4CL). The number of trials in each dataset and the data aggregation for each class is shown in Table 4.2.

We randomly divided the MF, PF, CF, and CRYS sets into two equal sized subsets, the training and the test sets. We used the training subsets to create the corresponding training DB_MF, DB_PF, DB_CF, and DB_CRYS datasets, and the tests subsets to create the independent test datasets. We designed our predictor based on the training datasets (using five-fold cross-validation protocol) and then we performed evaluation and comparison with the existing methods on the independent test datasets. The sequence identity between chains from the same class/outcome in the training and test sets is below 25%.

4.2.2 Features sources

Following the CRYSpred study, our method considers a comprehensive set of features generated using several information sources including the sequence and the sequence-derived isoelectric point, the encoding of amino acids in the sequence with several property-based indices (e.g., hydrophobicity and energy) from the AAIndex database (Kawashima et al., 2008), solvent accessibility predicted using Real-SPINE3 (Faraggi et al., 2009), disorder predicted using DISOPRED2 (Ward et al., 2004), and secondary structure predicted with PSIPRED 3.2 (Jones, 1999). The importance of the information derived directly from the protein chain, including the composition of certain amino acids, the isoelectric point, etc. for prediction of the crystallization success was demonstrated in numerous studies (Goh et al., 2004; Kantardjieff & Rupp, 2004; Kantardjieff et al., 2004; Overton & Barton, 2006; Smialowski et al., 2006; Chandonia et al., 2006; Chen et al., 2007; Slabinski et al., 2007a; Overton et al., 2008; Kurgan et al., 2009; Price et al., 2009; Kandaswamy et al., 2010). The usage of the energy and hydrophobicity of the residues is motivated by the work in (Goh et al., 2004; Overton & Barton, 2006; Chen et al., 2007; Overton et al., 2008; Kurgan et al., 2009; Price et al., 2009; Kandaswamy et al., 2010; Babnigg & Joachimiak, 2010). The predicted secondary structure, disorder, and solvent accessibility were found to be useful to predict propensity of the crystallization in (Chandonia et al., 2006; Slabinski et al., 2007a; Mizianty & Kurgan, 2009; Price et al., 2009; Kandaswamy et al., 2010). We note that we also use the above information to predict the propensity of the material production and purification, which is one of the novel aspects of this study. We considered 64 hydrophobicity- and energy-based indices from the AAIndex1 database and the side-chain entropy (Creamer, 2000) that was found useful in (Price et al., 2009); we disregarded amino acid indices related to the solvent accessibility and secondary structure, as we already include these predictions.

4.3 PPCpred

4.3.1 Considered features

We combine information based on the AA indices, predicted secondary structure and disorder with the predicted solvent accessibility by computing the values separately for the exposed and buried residues; we define buried residues as the residues for which

the predicted relative solvent accessibility is below 25%; otherwise a given residue is assumed to be solvent exposed. In total, we generated 817 features, which include 60 features based on AA composition (20 standard + 40 for exposed or buried residues only); pI; 704 features based on AA indices (average over an entire protein chain and for the exposed or buried residues only, and min and max values over sliding windows of sizes 5, 10, 15, and 20); 13 features based on predicted disorder, based both on predicted probabilities (average value over an entire chain and for the exposed/buried residues) and binary predictions (number, average and maximal length of predicted disordered segments, and content of exposed/buried disordered residues); 30 features based on the predicted secondary structure (average and maximal length of SS segments of a given SS type, average probability of each type of SS, and number of exposed/buried residues for each type of SS); and 9 features based on the predicted relative solvent accessibility (average value of predicted solvent accessibility, and distributions of exposed/buried residues with respect of the segment lengths they are in). The complete list of features together with a more detailed description is available in Appendix A.

4.3.2 Feature selection and the final design

We build a separate predictor for each training dataset using the same protocol. First we filtered features, starting with the feature with the highest absolute value of point biserial coefficient, by removing all co-correlated features (leaving only features for which all possible pairs of features have Pearson correlation less than 0.7), and removing feature which had low point biserial coefficient between two class labels. Then, for the remaining feature we applied the best first search using SVM classifier. We considered three SVM kernels: polynomial, RBF and sigmoid, and initially parameterized SVMs using top ten features with the highest point biserial coefficient. Finally the best model, in term of the MCC value, was selected for each dataset. These four models were then combined together to provide one final prediction. We considered two ways of merging classifiers, one (max-based) by selecting the class output/outcome from classifier which achieved the highest confidence score, and second by choosing the class/outcome based on the order of the steps in the crystallization protocol (order-based). In the second approach, we selected the outcome for the first classifier for which the predicted score was higher than selected threshold, and in case when all

probabilities are below the threshold, we selected the class with the highest probability. The order based approach turned out to provide better performance, as measured on the training dataset, and hence it was selected as the final model. The final architecture of our predictor is given in Figure 4.1.

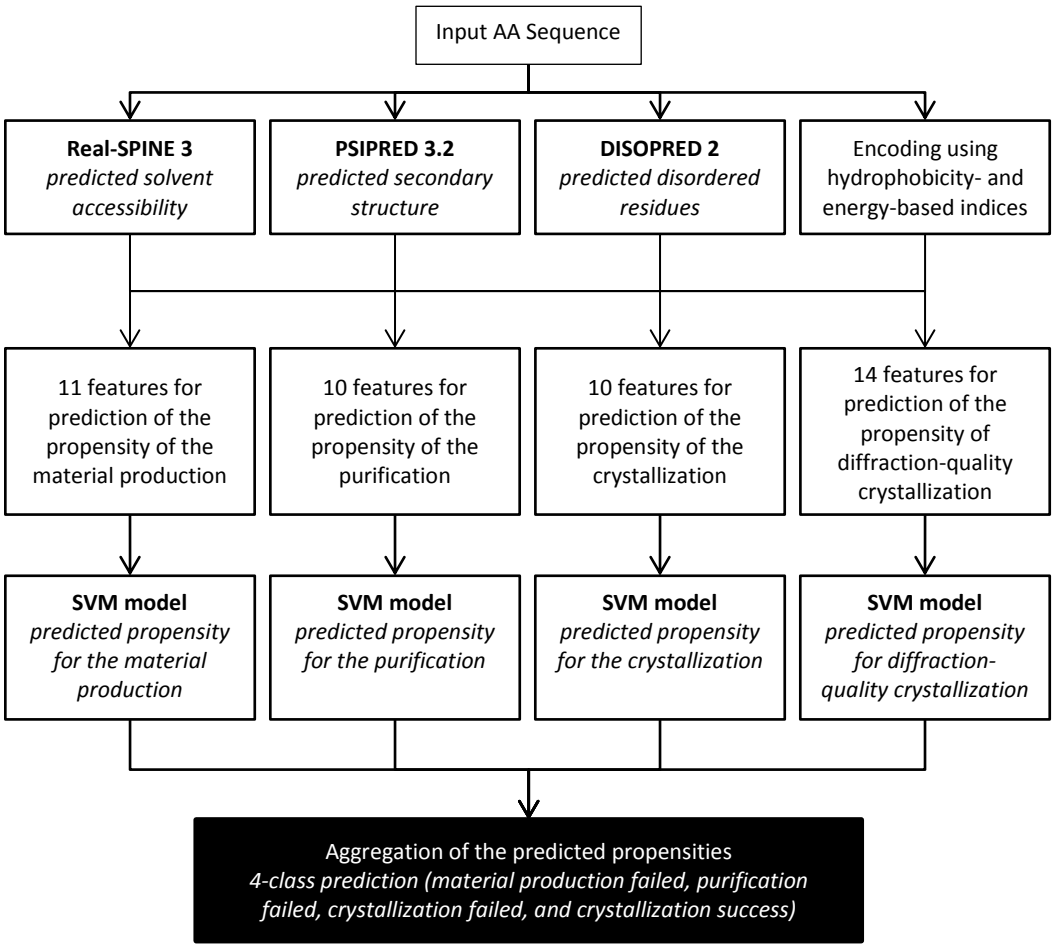


Figure 4.1: The overall architecture of the proposed PPCpred method.

4.4 Empirical evaluation

We compare PPCpred with previous approaches on the DB_CRYS dataset, as the previous methods can only distinguish between crystallizable and non-crystallizable proteins. For the remaining datasets, we compare our method to max based approach, selected SVM predictors, and to a commonly used BLAST-based solution. This solution uses PSI-Blast (Altschul et al., 1997) for each protein from test dataset to find the most

similar protein in the training dataset, and assign its class label/outcome to the test protein. The results on the four datasets are presented in Table 4.3.

Table 4.3: Summary of results for the prediction of the individual outcomes from crystallization pipelines. Predictions are based on following tables: the prediction of propensity of diffraction-quality crystallization success is based on the DB_CRYS test dataset, the prediction of the propensity of the material production failure is based on the DB_MF test set, the prediction of the propensity of the purification failure is based on the DB_PF test set, and the prediction of the propensity of the crystallization failure is based on the DB_CF test set. PPCpred is compared against results of OBScore, XtalPred, ParCrys, CRYSTALP2, MetaPPCP, and SVMCrys on the DB_CRYS dataset, and against the maximum-based aggregation method (max-based) and the BLAST-based predictor on the four datasets. The methods are sorted in the descending order based on their MCC scores, and the highest values for each quality index and dataset are shown in bold. The BLAST and SVMCrys provide only binary prediction and thus we could not compute their AUC. Results of tests of significance of the differences in MCC and ACC between PPCpred and the other methods are given in the "sig" columns. The tests compare values over 100 bootstrapping repetitions. The + and – mean that PPCpred is statistically significantly better / worse with p <0.01, and = means that results are not significantly different.

Test dataset (prediction target)	Method	MCC		Accuracy		SPEC	SENS	AUC
		value	sig	value	sig			
DB_CRYS (propensity of the diffraction-quality crystallization success)	**PPCpred**	**0.471**		**76.8**		84.8	61.2	0.789
	max-based	0.467	+	76.1	+	81.6	65.3	**0.793**
	SVM_POLY	0.398	+	74.6	+	**88.1**	47.9	0.779
	XtalPred	0.278	+	63.9	+	62.3	67.0	0.683
	SVMCrys	0.213	+	56.3	+	46.7	75.2	N/A
	CRYSTALP2	0.195	+	55.3	+	45.7	74.4	0.648
	MetaPPCP	0.195	+	59.9	+	59.0	61.7	0.620
	BLAST-based	0.188	+	65.6	+	79.5	38.0	N/A
	OBScore	0.124	+	47.8	+	31.4	**80.3**	0.572
	ParCrys	0.108	+	47.5	+	31.8	78.6	0.561
DB_MF (propensity of the material production failure)	**PPCpred**	**0.462**		**75.0**		**69.2**	78.0	0.755
	SVM_RBF	0.423	+	74.6	+	56.1	84.5	**0.791**
	max-based	0.339	+	71.6	+	45.4	**85.5**	0.621
	BLAST-based	0.014	+	55.4	+	35.3	66.0	N/A
DB_PF (propensity of the purification failure)	**PPCpred**	**0.324**		72.0		**50.1**	81.6	0.697
	SVM_POLY	0.290	+	**73.2**	-	30.8	**91.8**	**0.741**
	max-based	0.246	+	70.8	+	34.4	86.9	0.609
	BLAST-based	0.102	+	60.0	+	43.2	67.4	N/A
DB_CF (propensity of the crystallization failure)	max-based	**0.461**	–	76.9	–	70.5	79.2	0.813
	PPCpred	0.457		76.6		**70.8**	78.7	0.811
	SVM_POLY	0.346	+	**77.0**	=	40.1	**90.0**	**0.814**
	BLAST-based	0.060	+	60.9	+	37.0	69.4	N/A

The results on the DB_CRYS dataset shows that PPCpred outperforms the existing solutions in both the binary prediction (based on the MCC and accuracy) and the real-valued propensities (based on the AUC values). Statistical tests revealed that that the improvements in MCC and ACC offered by PPCpred are statistically significant. The best existing predictor is XtalPred, which is likely due to the usage of the sequence alignment against the PDB and nr databases, followed by SMVCrys and our MetaPPCP. The PPCpred improves over the SVM_POLY method, which demonstrates that aggregation of the results from the four SVMs is helpful. The order based selection used in PPCpred outperforms max-based method in binary prediction but the magnitude, although the difference is statistically significant, is small. The predictions from PPCpred are characterized by high specificity (high success rate among the non-crystallizable proteins) at about 85%. This means that we relatively rarely mispredict these chains to be crystallizable, which would save resources to solve other proteins.

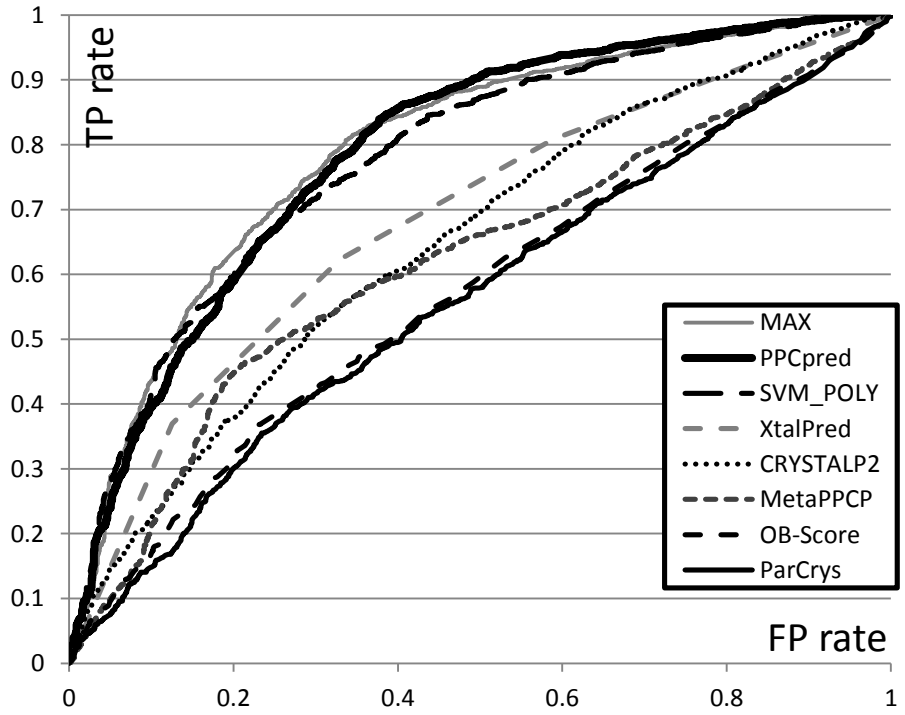


Figure 4.2: The ROC curves for the DB_CRYS dataset.
The ROC curves for the considered crystallization propensity predictors computed for the DB_CRYS test dataset.

The ROC curves of the considered predictors, except for the BLAST and SVMCrys that provide only the binary predictions, are shown in Figure 4.2. PPCpred outperforms the other solutions for TP-rates > 0.85 and FP-rates > 0.38, while the maximum-based aggregation works better for smaller TP- and FP-rates. This demonstrates that PPCpred is particularly useful when the user requires high TP-rates, i.e., the number of false negatives (crystallizable chains predicted are non-crystallizable) is low. In this case, PPCpred would relatively rarely mispredict chains that can be successfully solved, which would protect against abandoning solvable targets. The high TP-rate comes as a trade-off for higher FP-rate (higher rate of predicting non-crystallizable chains as crystallizable), which means that PPCpred would more often mistakenly advise to crystallize a difficult target, which consequently would waste resources.

The results for each individual target outcome of the PPCpred, BLAST-based predictors, our four SVM-based predictors of the material production, purification, crystallization, and diffraction-quality crystallization (SVM_POLY and SVM_RBF), and the maximum-based method for combining the four SVMs predictors are summarized in Table 4.3. Using the MCC measure, PPCpred significantly outperforms the other methods for the binary prediction of the material production, purification, and diffraction-quality crystallization, and it provides comparable predictive quality with the maximum-based aggregator for the prediction of the crystallization, i.e., the maximum-based aggregator provides an improvement with small magnitude that is statistically significant. PPCpred also provides well balanced values of the sensitivity and specificity. We note that our method provides reasonably high values of MCC, between 0.32 and 0.47, which indicate that it provides useful outputs.

We note that although the existing predictors achieve positive MCC values, they are generally lower than the values reported in the original publications on their test datasets. A possible explanation for that is that our annotation is somehow different and that the existing models were trained on relatively old trials for which crystallization experiments were performed before 2006. To test the latter hypothesis, we sorted the trials based on their date of the last activity to investigate whether the predictive quality varies with this timestamp, see Figure 4.3. The values of both MCC and AUC are lower for the more recent trials for majority of the methods, except for the PPCpred, XtalPred,

and the BLAST-based predictor. This confirms that the likely reason for the overall relatively low performance of the ParCrys, OBScore, CRYSTALP2, MetaPPCP, and SVMCrys is the fact that they utilize older training data. We note that XtalPred was updated in mid-2007 and it uses sequence alignment against recent contents of the PDB and nr databases, which helps to keep its predictions more up to date. This finding suggests that the advances in the crystallization protocols may render older predictors absolute, which motivates development of new, up-to-date methods.

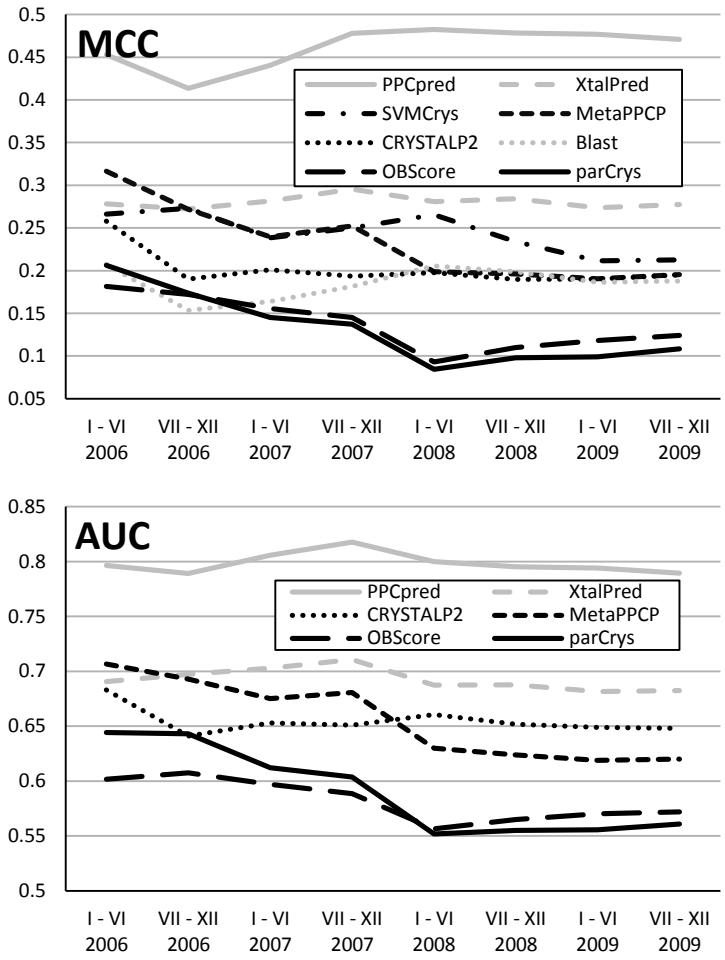


Figure 4.3: Comparison of results for prediction of crystallization propensity over time.
The MCC (top panel) and AUC (bottom panel) values obtained by the considered crystallization propensity predictors with respect with the date of the test trials (x-axis) from the DB_CRYS test dataset. BLAST and SVMCrys provide only binary prediction; their AUC cannot be computed.

The evaluation for the 4-class predictions on the DB_4CL test dataset is shown in Table 4.4. The output of the predictor indicates whether a given chain will provide high-quality crystal, will fail to crystallize, or whether the purification or material production will fail. The methods are evaluated using the overall accuracy (fraction of the correctly predicted chains) and mean MCC (over the four MCC values computed for each class/outcome). Only the PPCpred, the alignment based predictor, and the maximum-based aggregation method can be compared – the other methods predict only the outcome of crystallization. The overall accuracy of PPCpred equals 55.6%, which is higher by 5 and 21% than the accuracy of the other two solutions.

Table 4.4: Summary of results for the prediction of the all outcomes from crystallization pipelines. The results for predictions of failure in material production, failure in purification, failure in crystallization, and success in the generation of the diffraction-quality crystals, on the DB_4CL test dataset. The proposed PPCpred is compared against the maximum-based aggregation method (*max-based*), and the BLAST-based predictor. The methods are sorted in the descending order based on their MCC scores, and the highest values for each quality index and dataset are shown in bold. Results of tests of significance of the differences in mean MCC and ACC between PPCpred and the other methods are given in the "sig" columns. The tests compare values over 100 bootstrapping repetitions. The + and – mean that PPCpred is statistically significantly better / worse with p <0.01, and = means that results are not significantly different.

Method	mean MCC		Accuracy	
	value	sig	value	sig
PPCpred	**0.353**		**55.6**	
max-based	0.294	+	49.0	+
BLAST-based	0.041	+	31.1	+

4.4.1 Factors related to crystallization steps

The features selected for each classifier used in PPCpred are summarized in Table 4.5. A more detailed description of these features is presented in Appendix B. These features utilize all considered information sources, including the energy and hydrophobicity-based indices, composition of certain amino acid types, the predicted disorder, secondary structure and solvent accessibility, and content of certain buried and exposed residues. This shows that the success/failure in the considered steps of the crystallization process depends on a combination of multiple factors. We observe the strong presence of information derived from the hydrophobicity indices, which agrees with the observations in (Goh et al., 2004; Overton & Barton, 2006; Chen et al., 2007; Overton et al., 2008; Kurgan et al., 2009; Babnigg & Joachimiak, 2010), and from energy

based indices which so far were only used in MCSG Z-score (Babnigg & Joachimiak, 2010). Importantly, our features demonstrate the importance of the influence of the hydrophobic or hydrophilic/with high or low free energy patches/segments in the protein chain on the success/failure on all considered steps in the crystallization protocol, i.e., several selected features for prediction of each of the four considered steps are based on the minimal or maximal hydrophobicity in a sliding window. Our features also suggest the importance of Cys residues for the prediction of the material production and diffraction-quality crystallization, and buried Cys for the prediction of purification. This agrees with the observations in (Slabinski et al., 2007a; Overton et al., 2008), but these studies investigated the Cys residues only in the context of the propensity for the diffraction-quality crystallization and did not consider the influence of the solvent accessibility. Another factor related to the crystallization success is the content of the buried His. This agrees with the conclusions in (Overton et al., 2008; Kurgan et al., 2009), but again these studies considered only the overall content of this amino acid type, without the impact of the solvent accessibility.

Table 4.5: Summary of the PPCpred features.
Features types selected for the prediction of the material production, purification, and crystallization.

Features types	Number of features selected for the prediction of			
	material production	purification	crystallization	diffraction quality crystallization
hydrophobicity index	2	2	5	5
energy-based index	4	0	2	3
composition of AAs	1	3	1	1
isoelectric point	0	1	0	0
solvent accessibility	3	4	1	3
disorder	1	0	1	1
secondary structure	0	0	0	1
considered AA types	Arg, Cys, Glu	Asn, Cys, Ser, Met	His	Cys, His, Ser

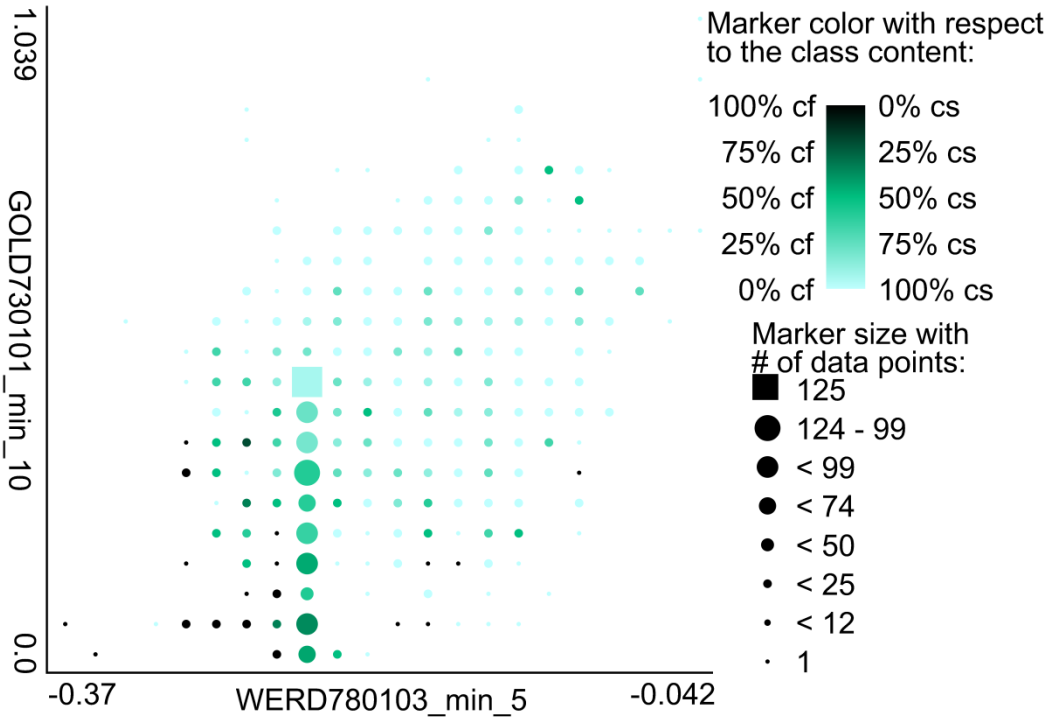


Figure 4.4: Scatter plot of values of a pair of features found to be good predictors for success of production of crystals from purified solution.
Size of the markers denotes the number of trials and color denotes their membership, green for the successful and black for the failed trials. The x-axis shows the minimal average value of the energy index (Wertz & Scheraga, 1978) in a window of 5 residues. The y-axis shows the minimal average value of the hydrophobicity index (Goldsack & Chalifoux, 1973) in a window of 10 residues. "cf" stands for crystallization failed, and "cs" stands for crystallization successful.

Figure 4.4 shows scatter plot of a representative pair of features that were selected for the prediction of the crystallization of a purified sample. The two features used to predict crystallization, GOLD730101_min_10 and WERD780103_min_5, are based on the minimal average values of the hydrophobicity (Goldsack & Chalifoux, 1973) and energy (specifically the energy of transfer in water of an isolated residue from a non-regular structure to the helical conformation) (Wertz & Scheraga, 1978) indices in the sliding windows of sizes 10 and 5, respectively. This means that the sequence segments with low hydrophobicity and transfer energy values are characteristic to chains that are difficult to crystallize. Importantly, combining these two features allows for improved separation between the successful and unsuccessful crystallization trials, i.e., trials for a given range of values of one index are further separated by the values of the other index.

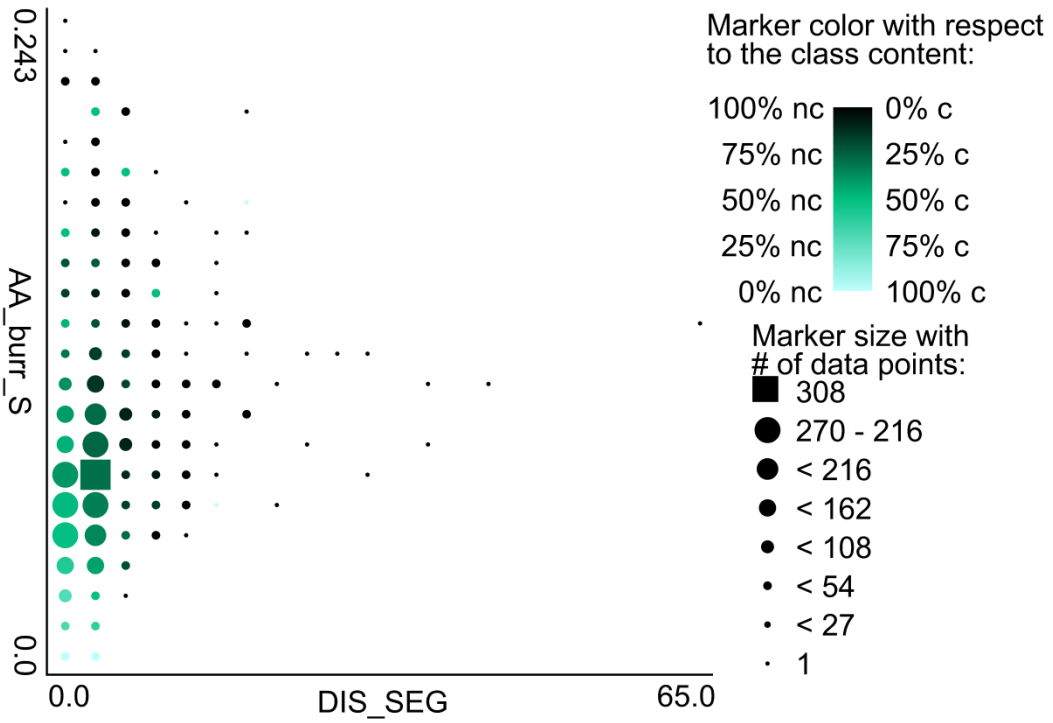


Figure 4.5: Scatter plot of values of a pair of features found to be good predictors for success of crystallization.
Size of the markers denotes the number of trials and color denotes their membership, green for the successful and black for the failed trials. The x-axis shows the number of the predicted disorder segments. The y-axis shows the content of the predicted buried Ser residues. "nc" stands for non crystallizable, and "c" stands for crystallizable.

The prediction of success of diffraction-quality crystallization from a given AA sequence (which includes passing all crystallization steps) is realized with the DIS_SEG and AA_bur_S features (see Figure 4.5), which quantify the number of the predicted disorder segments and the content of the predicted buried Ser, respectively. The content of Ser was shown to be important for the prediction of crystallization propensity in (Overton et al., 2008; Kurgan et al., 2009), but these studies investigated the overall Ser content, while we show that the (predicted) buried Ser provides strong discriminatory power. Similarly, while the content of the predicted disordered residues was used in several studies that predict the diffraction-quality crystallization (Slabinski et al., 2007a; Price et al., 2009), our analysis reveals the strong influence of the number of disordered segments. The plot shows that chains with larger number of disordered segments and larger number of buried Ser are more difficult to crystallize.

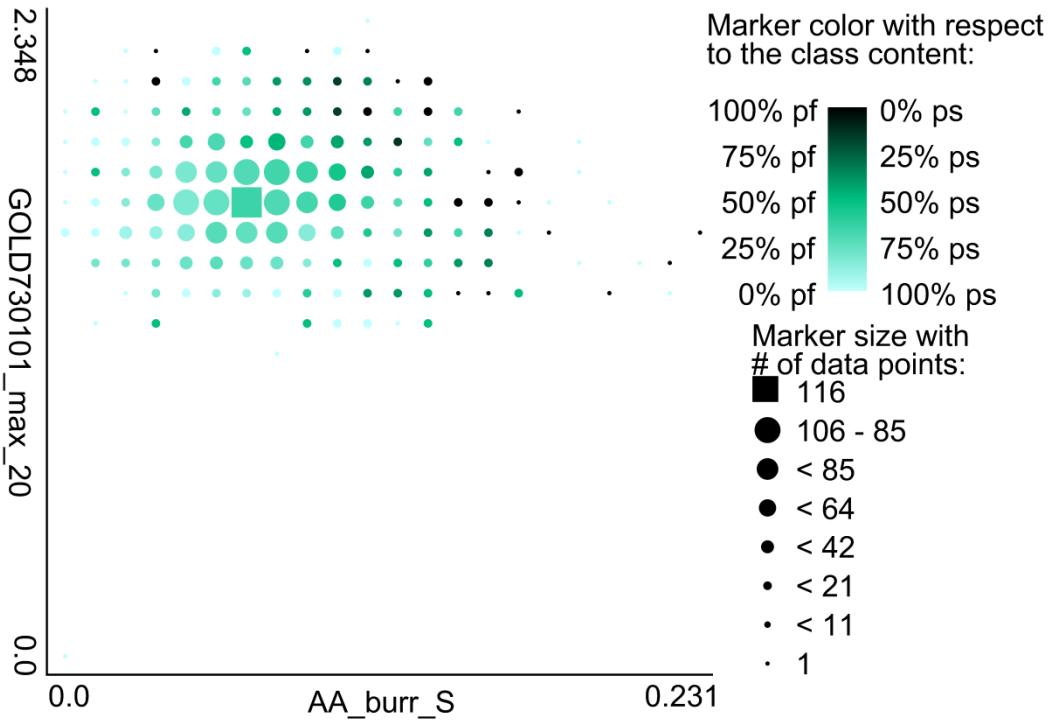


Figure 4.6: Scatter plot of values of a pair of features found to be good predictors for success of purification.
Size of the markers denotes the number of trials and color denotes their membership, green for the successful and black for the failed trials. The x-axis shows the content of the predicted buried Ser residues. The y-axis shows the maximal average value of the hydrophobicity index (Goldsack & Chalifoux, 1973) in a window of 20 residues. "pf" stands for purification failed, and "ps" stands for purification successful.

Finally, Figure 4.6 shows that chains with larger amount of buried Ser (AA_bur_S feature) and high hydrophobicity in a long sliding window (GOLD730101_max_20 feature, which denotes the maximal average values of the Goldsack-Chalifoux hydrophobicity index (Goldsack & Chalifoux, 1973) in sliding window of size 20) are more challenging to purify.

Overall, the factors that we identified are intuitive, physically reasonable, and they are well aligned with the "existing rules of thumb". Our main contributions are in providing additional details (e.g., related to solvent accessibility of selected residues types) and the fact that our model provides a novel way of balancing these factors to obtain good predictive performance.

4.5 Conclusions

We developed a first-of-its-kind in-silico method, PPCpred, which predicts the success/failure for four main steps in the protein crystallization protocols, including material production, purification, crystallization, and diffraction-quality crystallization. PPCpred is shown to significantly outperform alignment-based predictor as well as several modern crystallization propensity predictors. Our method provides the overall accuracy at 56% and average MCC at 0.35, which given current low success rates of the experimental protocols should provide useful input for the SG centers as well as individual crystallographers/biologists. In case of prediction of crystallization failure, our method points to the step in the crystallization pipeline which is likely to cause the failure. The webserver which implements PPCpred (available at: http://biomine.ece.ualberta.ca/PPCpred/) was released in June 2011. So far it was visited by 634 unique users from 260 cities in 48 countries, and was used 3,135 times to predict crystallization propensity of 12,814 protein chains; see Figure 4.7. This demonstrates that our method finds practical applications across the globe.

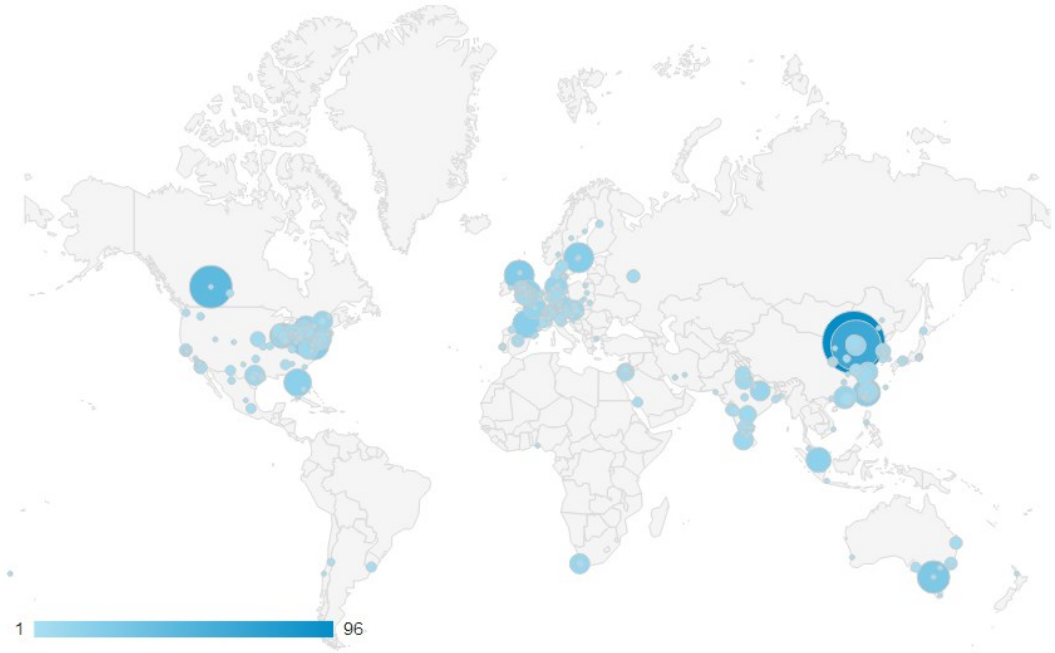


Figure 4.7: PPCpred webserver's usage demographics.
Map of the world showing PPCpred webserver's usage demographics. Each circle corresponds to city whereas its size and shade correspond to number of visits. Source: Google analytics.

Our analysis confirmed that predictive accuracy of crystallization predictors erodes over time (we hypothesized that in Chapter 3), most likely because of advances in crystallization protocols which enable crystallization of previously non-crystallizable targets. This motivates continuing research to keep the predictors up to date with respect to the crystallization protocols.

We also developed an improved protocol to annotate progress of protein chains along the crystallization process using the PepcDB, and we shows/confirm several interesting markers (based on the features included in our predictors) that influence the success/failure of the abovementioned steps. Although the predicted structural properties on input protein were found useful majority of features used in PPCpred were derived from AA sequence using AA indices. Generation of predictions of these structural properties is time consuming (in the order of several minutes per protein, depending on the protein size), which motivates investigation into building another crystallization propensity predictor that would be based only on fast-to-compute features (i.e. features based on information from AA sequence, or from predictions which could be generated quickly). Such predictor could be used on a proteomic scale, which is virtually impossible for PPCpred due to the considerably large computational cost.

Chapter 5

Analysis of attainable structural coverage based on predicted crystallization propensity

5.1 Introduction and motivation

As it was mentioned before, the X-ray crystallography is the main approach to experimentally solve protein 3D structure; it was used to solve almost 90% of known protein structures. Nevertheless, an interesting question is whether this approach is sufficient to solve all protein structures? Due to the large size of protein universe (set of all unique protein sequences) and the relatively low throughput of the X-ray crystallography, a likely answer is that this is not possible using crystallography alone. However, this question is more interesting if we couple X-ray crystallography with homology modeling. Here, crystallography would be used to solve a relatively small subset of proteins, which would then be used as a source of templates for comparative modeling algorithms. This approach is already applied by SG centers which focus on the protein-family oriented target selection, as they try to crystallize at least one representative from each protein family (Terwilliger et al., 1998). To answer the question about attainable coverage of X-ray crystallography and homology modeling, a large scale study of protein crystallization propensity is required. The propensities can be estimated with the use of machine learning algorithms, as we showed in the previous chapters. Using UniProt database (UniProt Consortium, 2012) we downloaded a current snapshot of protein universe, which consist of around ten million protein sequences from all currently fully sequenced proteomes, i.e., the sets of proteins thought to be expressed by an organism whose genome has been completely sequenced, as defined in UniProt. We extended our analysis of these proteomes to characterize structural

coverage of all functional and localization annotations available in Gene Ontology (GO) (Ashburner et al., 2000). We also selected a representative subset of these proteins by clustering the protein space at a 30% sequence identity threshold, a threshold which enables accurate homology modeling (Baker & Sali, 2001; Nair et al., 2009; Gront et al., 2012), and then selecting a representative protein from each cluster with the highest predicted crystallization propensity. This clustering approximates the use of homology modeling.

The study of crystallization propensity is important for several reasons. First, it may be used to characterize differences in crystallization propensity between different organisms and superkingdoms of life. Second, it could provide motivation for improving homology modeling and X-ray crystallography, as well as for the development of alternative methods for protein structure determination. Third, using the predicted crystallization propensities researchers can investigate a given organisms or functional annotation and create a list of the most feasible targets which are needed to increase structural coverage for a given organism/annotation. Finally, it will provide the answer to the question on how many protein structures can be solved using current X-ray crystallography protocols combined with homology modeling.

A method to perform such large-scale analysis must be both accurate and computationally efficient to handle the large protein dataset. Although our PPCpred predictor, which was introduced in the previous chapter, is accurate, it is not runtime-wise efficient. Its average runtime ranges between 5 and 15 minutes per protein, which would require almost 100 years to perform predictions for our analysis (assuming 5 minutes per-protein on a single thread). To this end, we developed an accurate and runtime-efficient method for Fast DEtermination of Targets' Eligibility for CrysTalization (fDETECT), which is capable of performing large-scale prediction of protein crystallization propensity.

In this chapter, we first introduce datasets used to develop and evaluate the new method and to perform the analysis. Next, we describe the development and empirical evaluation of the fDETECT predictor. Finally, we discuss the analysis, present our findings, and draw the corresponding conclusions.

5.2 Materials

5.2.1 Datasets

We used the training and test datasets that were previously used to design and evaluate PPCpred (Mizianty & Kurgan, 2011) to design the new crystallization propensity predictor. To allow for an accurate prediction of wild type protein sequences from UniProt we removed affinity tags (Waugh, 2005) from both training and test datasets. We note that the sequence identity between chains from the same class in the training and test sets is below 25%. Since we are interested in the prediction of crystallization propensity, we limited the class labels to non-crystallizable (which includes proteins which failed in material production, purification, and crystallization) and crystallizable.

To further evaluate our design, we used protein structures solved by X-ray crystallography with resolution no lower than 3.5 Å , which were deposited in the PDB (Berman et al., 2000) between January 1st 1993 and December 31st 2012. We filtered all redundant chains leaving only one representative chain from the structure with the highest resolution. The PDB datasets consist of 50,138 non-redundant chains collected from 44,671 X-ray structures of proteins.

Table 5.1: Summary of the datasets used to design and evaluate fDETECT and to perform structural coverage analysis.

Dataset	Number of proteins	Note
Training	3,587	Used to design fDETECT. Contains 1,204 crystallizable and 2,383 non-crystallizable proteins.
Test	3,584	Used to evaluate and compare fDETECT with existing predictors. Contains 1,204 crystallizable and 2,380 non-crystallizable proteins.
PDB	50,138	Non-redundant PDB chains from 44,671 structures with resolution no higher than 3.5 Å and deposited between Jan 1st 1993 and Dec 31st 2012
UniProt	9,586,243	1,953 fully sequenced proteomes proteomes (106 archaea, 1,043 bacterias, 265 eukaryotes and 539 viruses) from release 2012_07 of UniProt. Includes 8,652,940 non-redundant proteins.

The UniProt database consist of 9,586,243 proteins (8,652,940 non-redundant) from 1953 fully sequenced proteomes (106 archaea, 1043 bacteria, 265 eukaryotes and 539 viruses) collected from release 2012_07 of UniProt. The proteomes were assigned to their taxonomic lineage based on the NCBI BioSystems database (Geer et al., 2010). We also annotated proteins with Gene Ontology (Ashburner et al., 2000) terms where available. The considered datasets are summarized in Table 5.1.

5.2.2 Clustering and homology modeling

In order to investigate structural coverage of the current snapshot of the protein universe achievable through the combined use of X-ray crystallization and homology modeling (Gront et al., 2012), we clustered UniProt_DB using the UClust algorithm (Edgar, 2010). UClust could not process sequences longer than 10,000 AAs and thus they were removed from the analysis. We choose various thresholds of proteins identities, starting with the current threshold for homology modeling of 30% (Baker & Sali, 2001; Nair et al., 2009; Gront et al., 2012), i.e., current homology modeling provides good quality predictions in case of sequence similarity of at least 30%, and a few lower thresholds (20, 21, 23, and 25%), for which we compute coverage to estimate the magnitude of increased coverage due to improved quality of homology modeling. To analyze crystallization propensity for individual proteomes, we clustered each proteome separately at 30% sequence identity cut-off. We considered a given cluster to be solved at a given crystallization propensity score if at least one protein in the cluster has a score equal or above the given score, i.e. we assume that the remaining proteins from that cluster could be solved by homology modeling. Moreover, we calculate the coverage values over these clusters, i.e., the reported values are the number of solved clusters divided by the total number of clusters in a given analysis.

Finally, to estimate the current structural coverage we used the USearch algorithm (Edgar, 2010) to find all proteins from UniProt that have at least one target in PDB which covers no less than 90% of the query protein at no less than 30% sequence identity. As above, we assume that a given cluster could be solved by homology modeling if at least one of its members has such defined target in PDB i.e., a template structure for homology modeling is already available in the PDB.

5.3 Proposed approach

5.3.1 fDETECT

While designing our new method we expected that it will provide lower or comparable predictive quality when compared to PPCpred. We focused on obtaining good computational efficiency, so that the new method can be applied it on the proteomic scale. We also aimed to assure that fDETECT outperforms the other, older crystallization propensity predictors.

Features

PPCpred, although accurate, requires calculation of evolutionary information (so called PSSM profiles that are generated using PSI-Blast) to generate features concerning predicted secondary structure, disorder and relative solvent accessibility. Generation of the PSSM profiles takes at least a few minutes for each protein and this time is longer for larger proteins. To this end, we developed a fast and accurate algorithm which calculates its features only directly from a protein sequence.

We computed and evaluated total of 1,283 features/inputs for fDETECT. For the complete list of features along with their detailed description see Appendix C. The features are divided into following five groups:

- **420 amino acid based features.** These features include amino acid (AA) and dipeptides compositions.
- **336 amino acid group based features.** Features based on division of AAs into groups characterized by specific physicochemical properties, see Table 5.2. The twenty AAs are divided into three groups for each of the seven different AA characteristics representing the main clusters of the AA indices of Tomii and Kanehisaas (Tomii & Kanehisa, 1996) that were presented in (Dubchak et al., 1999).
- **448 amino acid index based features.** These features utilize per AA values of 64 hydrophobicity and energy based indices collected from the AAIndex database (Kawashima et al., 2008). The same indices that were used in PPCpred.
- **4 protein's properties based features.** Features based on physicochemical properties of proteins including: pI, aliphatic index, instability index, and net charge.

- **75 disorder and complexity predictions based features.** These features are computed from the predictions of disordered residues performed with IUpred (Dosztányi et al., 2005), which include predictions of both Short (IUpred_S) and Long (IUpred_L) disorder segments, and based on assignment of sequence complexity utilizing the SEG algorithm (Wootton & Federhen, 1993).

Table 5.2: Division of amino acids into groups based on their physicochemical and structural properties.

Characteristic	AA groups		
Hydrophobicity	Polar R, K, E, D, Q, N	Neutral G, A, S, T, P, H, Y	Hydrophobicity C, L, V, I, M, F, W
Normalized van der Waals volume	[0 – 2.78] G, A, S, T, P, D	[2.95-4.0] N, V, E, Q, I, L	[4.03 – 8.08] M, H, K, F, R, Y, W
Polarity	[4.9 – 6.2] L, I, F, W, C, M, V, Y	[8.0 – 9.2] P, A, T, G, S	[10.4 – 13.0] H, Q, R, K, N, E, D
Polarizability	[0 – 1.08] G, A, S, D, T	[0.128 – 0.186] C, P, N, V, E, Q, I, L	[0.219 – 0.409] K, M, H, F, R, Y, W
Charge	Positive K, R	Neutral A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	Negative D, E
Secondary structure	Helix E, A, L, M, Q, K, R, H	Strand V, I, Y, C, W, F, T	Coil G, N, P, S, D
Solvent accessibility	Buried A, L, F, C, G, I, V, W	Exposed R, K, Q, E, N, D	Intermediate M, P, S, T, H, Y

Feature selection and parameterization of classifiers

Features selection was performed in two steps. First, we filtered irrelevant (to the prediction of crystallization propensity) and redundant (cross-correlated) features. Second, we performed a wrapper-based feature selection utilizing the remaining features.

In the first step, we filtered the set of all considered features to include the features that are correlated to the crystallization outcome and to remove cross-correlated features. First, we removed all features with the biserial correlation with the class label (crystallization outcome) below the value of twice the average biserial correlation over the entire set of the considered features. Second, the removal of the cross-correlated features produced a set of features for which all possible pairs of features have Pearson correlations coefficient below 0.7 and 0.3 for the features set used in the later steps of the feature selection (58 features) and the features used to parameterize our predictive modle (11 features), respectively.

In the second step, we considered three types of classifiers: SVM (Vapnik, 1995), Logistic regression with ridge estimator (Cessie & van Houwelingen, 1992), and normalized Gaussian RBF network (Bugmann, 1998). We used LIBSVM (Chang & Lin, 2011) implementation for the SVM, and Weka (Hall et al., 2009) implementation of the latter two classifiers. For the SVM classifiers we considered linear, polynomial, RBF and sigmoid kernels. Each classifier (and kernel in case of SVM) was parameterized using eleven features selected in the first step of the feature selection to maximize AUC values. Next, we performed feature selection using the 58 features for each setup, i.e., each parameterized classifier including SVMs with different kernel types. Starting with the feature with the highest biserial correlation, we kept adding remaining features, which were ranked according to their biserial correlations, into the selected set of features if the addition of a given feature improves the AUC score.

The entire design process (parameterization and feature selection) was performed using five-fold cross-validation on the training dataset. The correlations and AUC values were computed as averages over the five training folds. We chose to optimize the AUC score as it evaluates predicted probability, which provides more information compared with the binary outcomes.

Evaluation and selection of the best design

We evaluated all considered setups based on their predictive quality and speed; the results are presented in Table 5.3. We note that the difference between the best and the worst AUC is relatively small, at about 0.04. We have selected logistic classifier with

11 features, which ranked as the second best in terms of AUC, as our selected design. This was motivated by the fact that this design offers three orders of magnitude faster runtime than the setup with the highest AUC, whereas the AUC score is worse only by a small margin of 0.001.

Table 5.3: Evaluation of the proposed designs based on five-fold cross-validation on the training dataset. Results are sorted according to AUC score, the best value for each measure is shown in bold font, and the selected design is highlighted in grey.

Method	Runtime per protein [s]	# of feat.	Accuracy [%]		MCC		Specificity [%]		Sensitivity [%]		AUC	
			avg	std	avg	std	avg	std	avg	std	avg	std
LIBSVM_RBF	0.217	11	72.7	0.5	.339	.009	88.8	0.9	40.5	1.1	**.772**	.004
LOGISTIC	0.002	11	72.5	0.6	.327	.009	90.8	0.7	35.9	0.8	.771	.004
LIBSVM_LIN	0.163	11	72.4	0.6	.334	.007	87.5	1.0	41.9	1.3	.769	.004
RBF NETWORK	0.003	11	**72.9**	0.7	**.359**	.011	85.1	1.1	48.3	1.3	.769	.006
RBF NETWORK^{fs}	0.002	8	72.8	0.5	.349	.009	86.9	0.9	44.4	1.0	.762	.006
LOGISTIC^{fs}	**0.001**	7	72.1	0.6	.309	.008	**92.1**	0.7	31.7	1.3	.756	.005
LIBSVM_POLY	0.143	11	70.9	1.0	.311	.035	85.7	7.1	41.6	13.7	.755	.008
LIBSVM_LIN^{fs}	0.099	6	71.6	0.4	.303	.010	89.3	0.8	35.9	1.7	.753	.004
LIBSVM_SIG	0.251	11	71.1	1.2	.331	.031	82.9	7.0	47.6	12.6	.753	.009
LIBSVM_RBF^{fs}	0.186	8	69.9	1.1	.313	.040	80.2	8.9	49.3	15.7	.746	.012
LIBSVM_POLY^{fs}	0.110	8	67.6	3.8	.303	.041	75.0	15.0	**53.2**	21.2	.743	.014
LIBSVM_SIG^{fs}	0.206	6	67.0	4.5	.287	.041	75.6	16.0	50.1	20.7	.732	.008

^{fs} – designs with feature selection

5.3.2 Empirical evaluation

We evaluated fDETECT on the test dataset and compared it with OB-Score (Overton & Barton, 2006), XtalPred (Slabinski et al., 2007a, 2007b), CRYSTALP2 (Kurgan et al., 2009), MetaPPCP (Mizianty & Kurgan, 2009), SVMCrys (Kandaswamy et al., 2010), and PPCPred (Mizianty & Kurgan, 2011) methods. The results presented in Table 5.4 show that fDETECT obtained the highest AUC score, and the second best, behind PPCpred, accuracy and MCC. Most importantly, the empirical results reveals that fDETECT, while having similar prediction quality with the best performing PPCpred, is 6 orders of magnitude faster than PPCpred and the third best method, XtalPred. Our evaluation also shows that fDETECT is slightly slower than CRYSTALP2, but its accuracy, MCC, and AUC values are significantly higher by 14% points, 0.152, and .096, respectively. The measures of predictive quality were calculated as an average over 100 repetitions of randomly selected 50% of the data, and the differences between scores where

compared using the Student's paired t-test if distributions were normal; or with the Wilcoxon test otherwise. Distribution type was verified using the Anderson-Darling test.

Table 5.4: Comparison of fDETECT and other crystallization propensity predictors on the test dataset. Results are sorted according to AUC score, the best value for each measure is given in bold font. Results are reported as average (avg) and standard deviation (std) values for 100 repetitions over randomly selected 50% of the chains from the test dataset; +/- in the sig columns denote results that are statistically significant worse/better than the corresponding results of fDETECT.

Method	Runtime per protein [ms]			Accuracy [%]			MCC			Specificity [%]		Sensitivity [%]		AUC		
	avg	std	sig	avg	std	sig	avg	std	sig	avg	std	avg	std	avg	std	sig
fDETECT	0.8	0.0		70.6	0.8		.354	.017		75.8	0.8	60.3	1.5	**.754**	.009	
PPCpred	1529139	1438	+	**71.8**	0.8	-	**.361**	.017	-	**79.7**	0.8	56.0	1.5	.741	.009	+
XtalPred*	70624	1008	+	53.3	0.9	+	.248	.016	+	36.0	1.0	**87.6**	1.1	.665	.011	+
CRYSTALP2	**0.3**	0.0	-	56.6	0.8	+	.202	.015	+	48.5	0.9	72.6	1.3	.658	.010	+
SVMcrys	153	0.7	+	56.5	0.8	+	.223	.017	+	46.5	1.0	76.5	1.4	.615	.009	+
OBScore	64	0.2	+	47.2	0.9	+	.130	.017	+	29.3	1.0	82.7	1.1	.569	.010	+
ParCrys**	N/A	N/A	N/A	48.3	0.8	+	.105	.016	+	34.5	0.9	75.9	1.1	.557	.010	+

* XtalPred results were obtained from a webserver, the time estimation may be inaccurate

**ParCrys is available as webserver and we could not estimate its time efficiency

We also evaluated fDETECT on the PDB dataset. Interestingly, our analysis showed a trend between resolution of structures in this dataset and our predicted crystallization propensity scores. Namely, higher scores are correlated with higher, on average, structure resolution, see Figure 5.1. We obtained a good fit into these data with a third degree polynomial, which corresponds to inverted cubical ($1/r^3$) nature of crystal resolution.

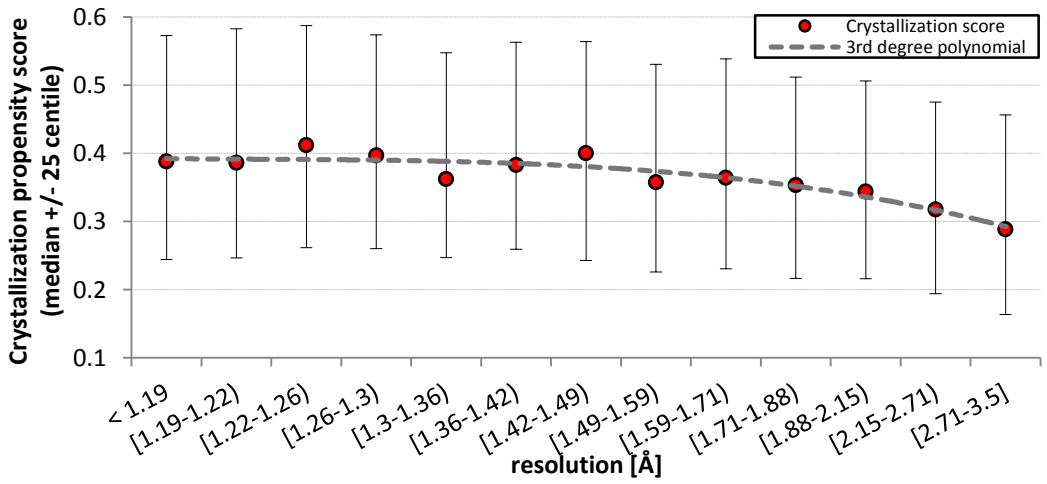


Figure 5.1: Trend between predicted crystallization propensity and crystal resolution.
The box plot shows 25th, 50th, and 75th percentile of scores for crystal structures with resolution from a given range. Ranges for resolutions were selected to reflect inverted cubical nature of crystal resolution. Dashed line represents fitted 3rd degree polynomial.

Our method was also validated to show that average crystallization propensity scores computed for chains with structures in the PDB (resolution < 3.5 Å) are higher than the average scores for all chains from the same organisms, see Figure 5.2. The positive relative difference, which is given on the *x*-axis, denotes that scores for PDB structures are higher than for all chains, which agrees with our expectation that on average crystallographers would solve more chains which were easier to crystallize. Moreover, we observed two trends: a) relative differences are lower for bacterial proteomes, which overall have high propensity for crystallization, and b) relative differences are high for eukaryotes, i.e., the already solved structures have substantially higher propensity for crystallization compared to the corresponding overall propensity; this means that the remaining to solve chains are going to be likely harder to crystallize.

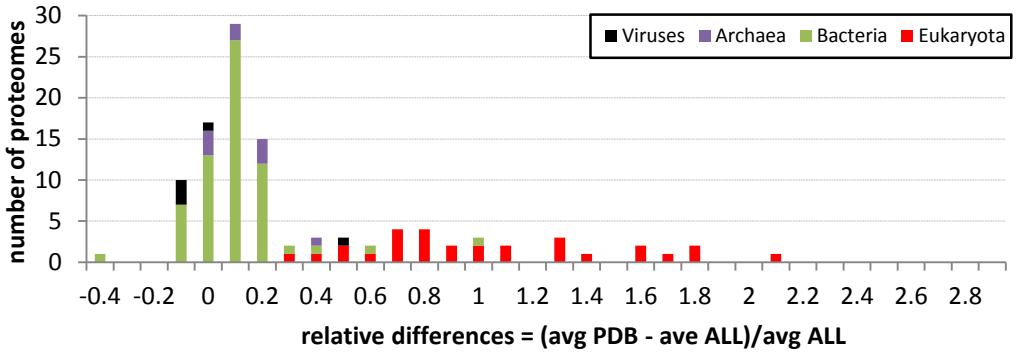


Figure 5.2: The relative difference in crystallization propensity scores between the proteins deposited in PDB and UniProt.
Graph compares the predicted crystallization propensity for all proteins from a given organism and the proteins from a given organism which were deposited in the PDB. We selected organisms with at least 20 chains deposited in the PDB.

5.3.3 Features related to crystallization

fDetect uses eleven features which are correlated with the crystallization propensity and which are non-redundant with each other. Three of these features are based on AA compositions, another three and two are based on energy and hydrophobicity based indices, respectively, and the remaining three correspond to the instability index, distance of the first AA of medium polarizability from the N-terminus, and fraction of long segments (15 AAs or longer) that are characterized by high amino acid complexity.

Figure 5.3 presents the box plots of values of the eleven features on the training dataset along with their biserial correlation (with the binary crystallization output).

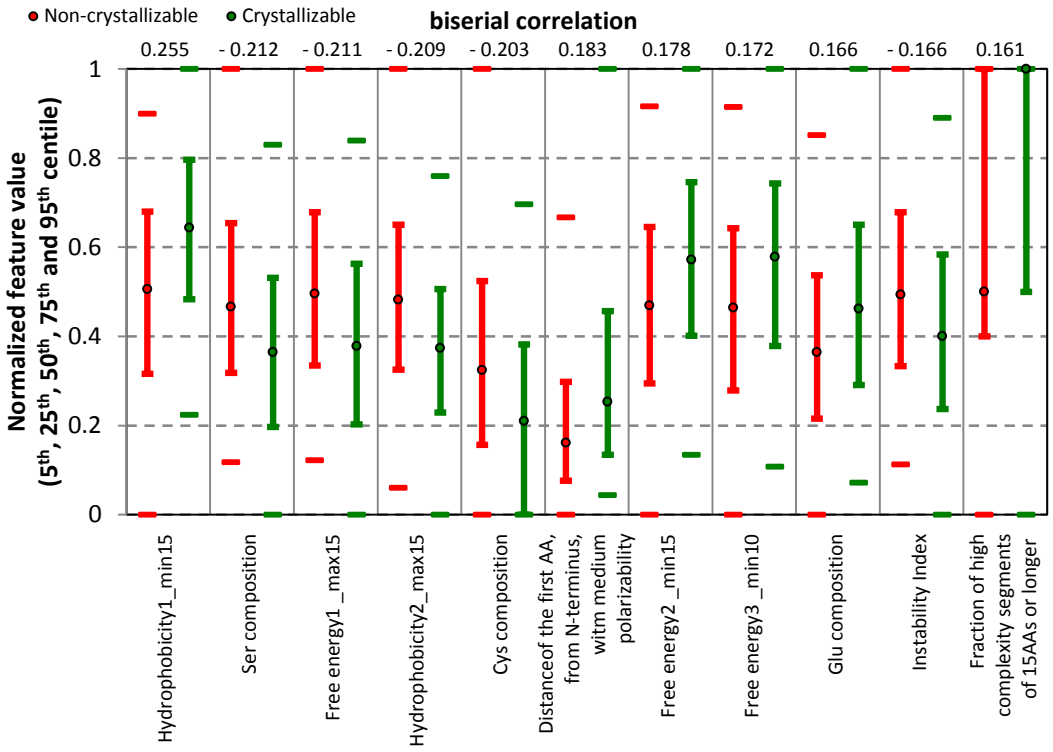


Figure 5.3: Distribution of normalized features' values on the training dataset.
Values were normalized using min max normalization where 5th and 95th centile values were selected as minimal and maximal values, respectively. The box plots show the median (50th centile, hollow circle), 25 and 75th centiles (whiskers) and the min and max values (dash markers). Features are sorted from left to right according to their absolute biserial correlation, shown at the top of the figure, in the descending order. Hydrophobicty1 refers to MANP780101 index, Hydrophobicty2 to CASG920101 index, Free energy1 to WERD780102 index, Free energy2 to RADA880103 index, and Free Energy3 to WERD780103 index. Indices were taken from the AAIndex database. min*X* and max*X* refer to minimal and maximal average value of a given index over possible segments of *X* neighboring residues in the sequence.

Hydrophobicity-based features (features which are based on MANP780101 "Average surrounding hydrophobicity" (Manavalan & Ponnuswamy, 1978) and CASG920101 "Hydrophobicity scale from native protein structures" (Casari & Sippl, 1992) indices) show that non-crystallizable proteins tend to have long segments characterized by a wider range of hydrophobicity (lower values for the minimum in the MANP780101-based feature and higher values for the maximum in the CASG920101-based feature), whereas crystallizable proteins tend to exclude long segments with either high or low hydrophobicity. The hydrophobicity of the protein chain has been linked with crystallization outcome in many studies (Goh et al., 2004; Overton & Barton, 2006; Chen et al., 2007; Overton et al., 2008; Kurgan et al., 2009; Price et al., 2009; Babnigg & Joachimiak, 2010; Overton et al., 2011), and two of these studies also investigated hydrophobicity in segments of a protein sequence (Babnigg & Joachimiak, 2010; Mizianty & Kurgan, 2011).

The distribution of values of the three free energy-based features (features which are based on WERD780102 “Free energy change of epsilon(i) to epsilon(ex)” (Wertz & Scheraga, 1978), RADA880103 “Transfer free energy from vap to chx” (Radzicka & Wolfenden, 1988), and WERD780103 “Free energy change of alpha(Ri) to alpha(Rh)” (Wertz & Scheraga, 1978) indices) show that the non-crystallizable proteins are more likely to include segments with higher and lower free energy change values, whereas crystallizable proteins consist of regions with medium free energy change values; this is similar to the observation related to hydrophobicity. The indices related to the free energy changes were also used to design PPCpred (Mizianty & Kurgan, 2011) and MCSG Z-score (Babnigg & Joachimiak, 2010).

Crystallizable proteins are shown to be enriched in Glu, whereas high content of Ser and Cys is characteristic to proteins which are hard to crystallize. This agrees with observations in (Babnigg & Joachimiak, 2010; Mizianty & Kurgan, 2011); the Glu content has been also used in (Price et al., 2009), whereas Ser and Cys contents were used in (Overton et al., 2008) and (Overton et al., 2008; Slabinski et al., 2007a), respectively.

Instability index, with higher values denoting instable proteins with shorter *in vivo* half-life, tends to be higher for the non-crystallizable proteins, which agrees with finding in (Slabinski et al., 2007a).

Crystallizable proteins are shown to have a large fraction of long high complexity segments predicted by the SEG algorithm (Wootton & Federhen, 1993). In fact, over 50% of crystallizable proteins have no low complexity segments. Low complexity regions were linked to disorder, with a general rule that inclusion of a larger number and longer low complexity regions implies higher content of disorder (Romero et al., 2001). Information about the predicted disorder was used to determine protein crystallizability in five previous studies (Oldfield et al., 2005; Slabinski et al., 2007a; Price et al., 2009; Mizianty & Kurgan, 2011, 2012). Also, the sequence complexity, more specifically number of low-complexity regions, was linked with crystallization in two other studies (Canaves et al., 2004; Chandonia et al., 2006).

Interestingly, it seems that the non-crystallizable proteins have AAs with medium polarizability (Cys, Pro, Asn, Val, Glu, Gln, Ile, Leu) closer to the N-terminus than the crystallizable targets. We hypothesize that this could be due to an interaction with affinity tags which are attached mostly to protein’s N-terminus.

Except of the last feature, the characteristics associated with the features utilized by fDETECT are well grounded in the literature and have been shown to be markers of crystallization outcomes. This study formulates a novel combination of these characteristics that can be calculated quickly and which offers competitive levels of predictive performance for the prediction of the crystallization propensity.

5.4 Attainable structural coverage analysis

Our analysis aims to reveal the attainable structural coverage that combines usage of homology modeling and current crystallization protocols that utilize target selection. We investigated the propensity for crystallization at the proteome and superkingdom levels for the current snapshot of the protein universe (8,652,940 non-redundant proteins encoded in 1,953 fully sequenced genomes). We also comprehensively estimated the structural coverage of GO annotations including the annotations of molecular functions, biological processes, and cellular components. Finally, we present results for two case studies: *Homo sapiens* proteome, and transmembrane proteins G-Protein Coupled Receptors (GPCRs). The *H.sapiens* study reveals extend of the current structural coverage and attempts to estimate how large this coverage could become when utilizing current crystallization protocols and homology modeling algorithms. The analysis of GPCRs confirms that this group of protein is difficult to crystallize and shows that our method can be used to point out to targets which are potentially more suitable for crystallization. We selected GPCRs as our case study as this is a very important family of membrane proteins involved in cellular signaling that encode roughly 21% of the genes of known function (Roth, 2005; Schwartz & Hubbell, 2008) and represent 50–60% of current drug targets (Lundstrom, 2009).

5.4.1 Characterization of crystallization propensity for proteomes

Using fDETECT we have calculated crystallization propensity of 1,486 proteomes (106 archaea, 1,041 bacteria, 265 eukaryotes and 37 bacterial and eukaryotic viruses) which, to assure statistically sound estimates, consist of at least 100 clusters at 30% sequence identity. We investigate overall crystallizability of the considered proteomes by analyzing their median values of the crystallization propensity predicted by fDETECT over protein clusters, where each cluster is represented by a protein with a maximal crystallization propensity score. Our analysis reduces sequence redundancy within

individual organisms, by clustering at 30% sequence identity, to assure that results are not affected by homology to other proteomes.

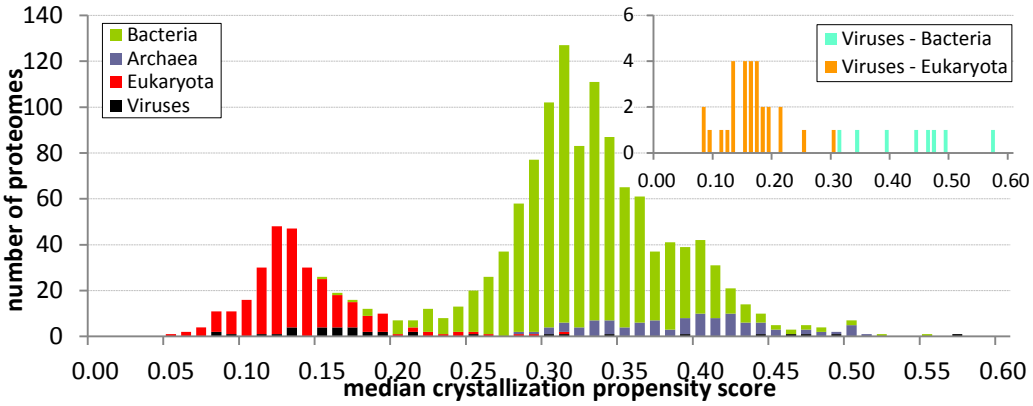


Figure 5.4: Crystallization propensity across all considered proteomes.
To assure statistically sound estimates we limited analysis to organisms with at least 100 clusters at 30% sequence identity. Clustering was done at the organism level to assure that results are not affected by homology to other proteomes. Median crystallization propensities were computed across all maximal scores per cluster. The inset divides viruses by superkingdoms of their host organisms. Viruses hosted by archaea all have proteomes below 100 clusters and hence were not included in the above graph.

Figure 5.4 reveals that the overall distribution of propensities is bimodal, suggesting that there are two clusters of proteomes: hard to crystallize (peak on the left), which includes most of eukaryotes and eukaryotic viruses; and easy to crystallize (peak on the right), which include most of bacteria, archaea, and bacterial and archaean viruses. Archaean proteomes, when compared with bacterias, are shifted further towards higher scores. Interestingly, this trend correlates with overall complexity of these organisms, where archea are the least complex, followed by bacteria and eukaryotic organisms.

5.4.2 Attainable structural coverage

We analyze attainable structural coverage using X-ray crystallography combined with homology modeling at the whole proteome level (% protein clusters that are above a certain value of crystallization propensity) for each considered organisms grouped per superkingdom. We assumed median score obtained for proteins for which structures are deposited in PDB (0.336) as our cut-off on crystallization propensity, i.e., proteins with scores above this cut-off are assumed to be solvable. Similarly to previous study, to assure statistically sound estimates we limited this analysis to a subset of proteomes that have at least 100 clusters at the 30% sequence identity. Results of this analysis are presented in Figure 5.5. The top three lines in this Figure show coverage that can be obtained when structures of all proteins with scores above a median score for proteins

in the PDB are determined and when the structures of related homologs are produced by homology modeling. To estimate the effects of potential improvements in homology modeling each protein was mapped to the clustered (at 30%, 25%, and 20%) UniProt set (combined set of all complete proteomes) and assigned the highest score in the corresponding cluster for this protein.

The "Random target selection" plot shows the attainable coverage when proteins with an average score are used for structure determination instead of proteins with the best scores. This corresponds to a more traditional way of selecting protein targets when crystallizability propensity score is not used to prioritize targets. As a reference point, the "PDB coverage" plot shows the structural coverage that can be achieved currently using the same 30% sequence identity cut-off based homology modeling with the current contents of the PDB as the modeling templates.

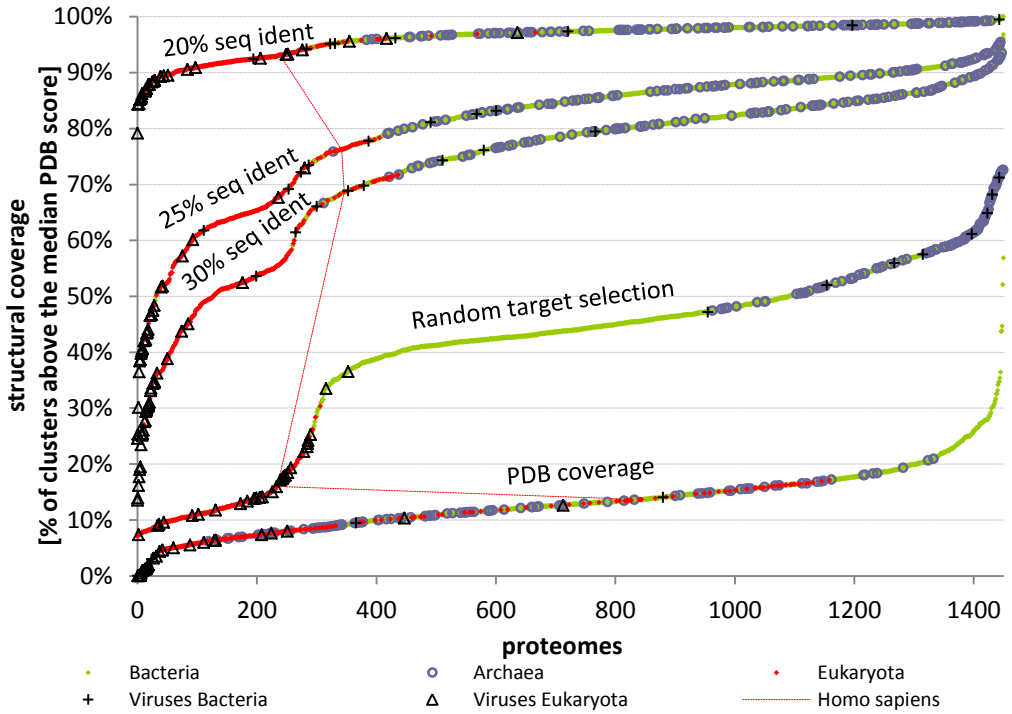


Figure 5.5: The current and attainable structural coverage per proteomes.
Current and attainable structural coverage is shown using points which are grouped into lines depending on specific criteria used. For each line (except "PDB coverage") proteomes are sorted based on their attainable structural coverage which is determined using a threshold on probabilities equal to the median score for PDB structures. The top three lines show the coverage when proteomes are clustered at different levels of sequence identity (20, 25, and 30%), the line titled "Random target selection" shows the structural coverage where instead of using a maximal score per cluster we are using an average score, and the line labeled "PDB coverage" refers to the actual coverage which could be obtained now by homology modeling using the existing structures in PDB as templates. To assure statistically sound estimates we limited analysis to proteomes with at least 100 clusters at 30% sequence identity.

The plot that includes a 30% sequence identity homology modeling cut-off shows that virtually all bacterial and archaeal organisms and bacterial viruses can be covered at above 70%. There is a visible decline in coverage between combined bacteria and archaea proteomes and the lower part of the figure (eukaryotic viruses and eukaryotes). Most of the eukaryotes and eukaryotic viruses have coverage below 70%.

Our analysis also shows that improving homology modeling, so the proteins with lower levels of identity could be modeled, would bring dramatic improvements. At the 20% sequence identity, structural coverage would include 90% of clusters (templates) for virtually all organisms, except for the few eukaryotic viruses where coverage would be over 85%. A less ambitious improvement in homology modeling (25% cut-off) would still lead to improvements in coverage by about 10%.

These results demonstrate that use of target selection and prioritization based on the crystallizability score allows to substantially, by 25 to 40%, improve structural coverage (measured as the increase between “Random target selection” and “30% seq ident” plots), except for few eukaryotic viruses for which coverage would register smaller, 5 to 25%, improvement.

5.4.3 Attainable structural coverage of protein families

Figure 5.6 shows attainable structural coverage (% clusters/templates per proteome that are above a certain value of the crystallization propensity score) for 1,734,048 protein families clustered at 30% sequence identity level and aggregated per superkingdom. To include the contribution of homology modeling, each protein was mapped to the clustered, at 30% sequence identity, UniProt (combined set of all complete proteomes) and we assigned the highest score in the corresponding cluster for this protein. This analysis includes also proteins from proteomes which were excluded from previous analyses due to the small sample size.

Results from this study agree with the results on characterization of propensity for crystallization at the proteome level. Figure 5.6 shows that crystallization propensity varies between superkingdoms, with the same order of difficulty in crystallization as we have shown above. When assuming cutoff at the median crystallization propensity score

for proteins in the PDB (0.336), the coverage ranges from 32% for eukaryotes to 70% for archaea, with the overall structural coverage over all superkingdoms at 43%.

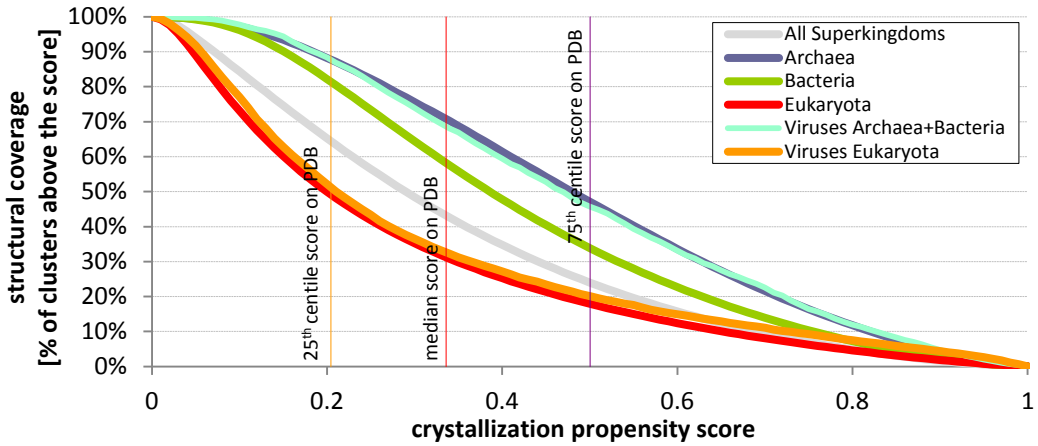


Figure 5.6: The attainable structural coverage of the current snapshot of protein universe.
The UniProt database was clustered at 30% sequence identity and each cluster is considered solved when the crystallization score is above a given cut-off. The lines show the structural coverage (number of solved clusters to all clusters in a given superkingdom) when the value on the x-axis is considered a cut-off on crystallization propensity score.

5.4.4 Attainable structural coverage of GO annotations

Important consideration is how many structures of proteins with different functions can be obtainable through experimental and computer-based approaches. To this end we measured the attainable structural coverage of various cellular functions, processes and components, as represented by the GO annotations, aggregated per superkingdom. To simulate the use of homology modeling each annotation, i.e., a set of protein with a given annotation, was mapped to the clusters from UniProt set, clustered at 30% sequence identity.

Assuming that only one structure is required to be solved per annotation and using the median cut-off with homology modeling at the 30% sequence identity cut-off, virtually all annotations can be covered by structures, see Figure 5.7.

However, if one wants to obtain structures for at least half of the clusters for a given annotation, the coverage varies considerably between superkingdoms with 79% for eukaryotes, 90% for bacteria, and 97% for archaea. Viruses show relatively poor coverage at only 6%. Moreover, assuming full structural coverage for each annotation (all representative clusters in each annotation are solved), the coverage is reduced significantly and approaches 0% for eukaryotes and viruses, 4% for bacteria, and 10% for

archaea. This shows that most of the annotations include proteins with structures that are currently difficult to solve.

We also investigated the structural coverage for each of the three types of GO annotations, including cellular components, molecular functions, and biological processes (see Appendix D). These per-annotation-type results follow the same trends as the overall results. Considering the solid lines (coverage based on the highest scoring single chain), all superkingdoms are grouped together, i.e., they have similar and high coverage. Considering the dashed lines (coverage based on solving 50% of families per annotation), the superkingdoms are well separated and in the same order irrespective of the annotation type, from the easiest to cover archaea and the hardest viruses.

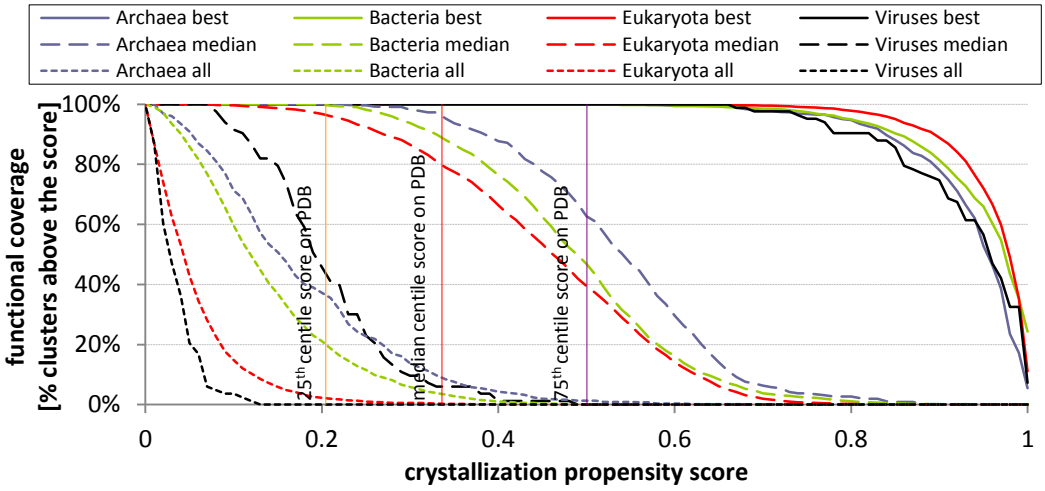


Figure 5.7: The attainable structural coverage of GO annotations.
The UniProt database was clustered at the 30% sequence identity and each cluster is considered solved when the crystallization score is above a given cutoff (*x*-axis). The lines show the functional coverage (fractions of solved GO annotations among all GO annotations in a given superkingdom) when the value on the *x*-axis is considered a cutoff for the crystallization propensity score. The thick solid lines represent coverage where a given annotation is assumed covered when one of its annotated clusters has score above the cut-off (at least one cluster for a given annotation can be solved). The dashed lines represent coverage where a given annotation is assumed covered when at least 50% of clusters in this annotation are covered; the dotted line is when all clusters in a given annotation are covered (the annotation is fully covered). To assure statistically sound estimates we limited analysis to the annotations with at least 20 clusters.

5.4.5 Analysis of human proteome

Obtaining structures or accurate homology models for human proteins is of high value and hence we performed an analysis of Homo sapiens proteome, which is shown in Figure 5.8. Results for functional GO annotations are presented in panel A, whereas panel B compares the actual and attainable structural coverage for GO annotations.

Moreover, dotted line in Figure 5.5 shows attainable structural coverage for all proteins encoded in the *Homo sapiens* proteome.

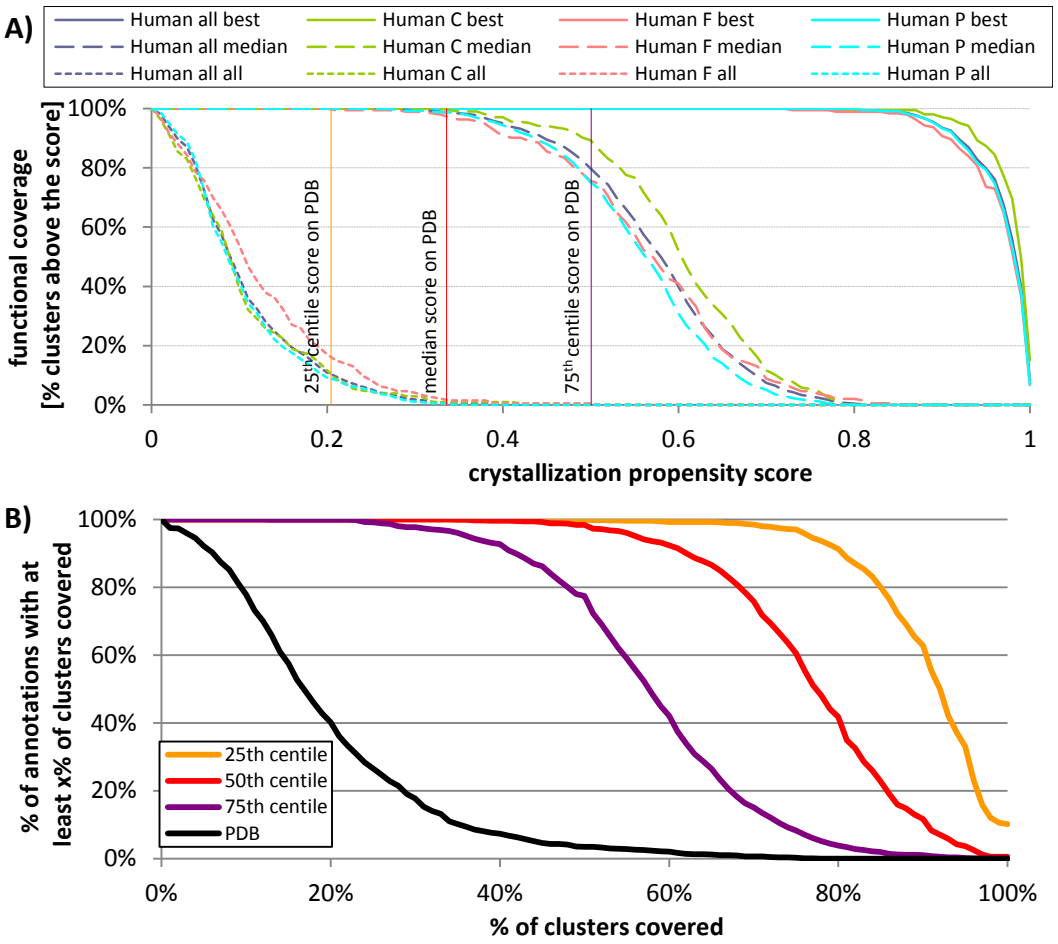


Figure 5.8: The current and attainable structural coverage of *Homo sapiens* GO annotations.
The Uniprot database was clustered at the 30% sequence identity and each cluster is considered solved when the crystallization score is above a given cutoff value, only clusters with proteins from the *Homo sapiens* proteome were used. Panel A shows the functional coverage (fraction of solved GO annotations among all GO annotations in a given category) when the value on the *x*-axis is considered a cut-off for the crystallization propensity score. The thick solid lines represent coverage where a given annotation is assumed covered when one of its annotated clusters has score above the cut-off (at least one cluster for a given annotation can be solved). The dashed lines represent coverage where a given annotation is assumed covered when at least 50% of clusters in this annotation are covered; the dotted line is when all clusters in a given annotation are covered (the annotation is fully covered). C, F, and P stand for Component, Functional and biological Process annotations, respectively. To assure statistically sound estimates we limited analysis to annotations with at least 20 clusters. Panel B shows current and attainable coverage of the annotated *Homo sapiens* proteome. The *y*-axis shows the % of annotations which have at least *x*% of clusters covered, where *x* is given on *x*-axis. Lines labeled as 25, 50 and 75 centile are the structural coverage when we assume cutoff for the crystallization propensity score equal to 25th, 50th and 75th centile, respectively, of scores for the PDB structures.

Given an objective to determine at least one structure per annotation based on the median cutoff of the crystallization score, one can structurally cover virtually all annotations across all annotation types (i.e., component, function and biological process

annotations) for the human proteome. The coverage drops slightly to an overall (across all annotation types) value of 99% if we want to cover at least half of the clusters for a given annotation to consider it structurally covered. These coverage values are higher than corresponding values for Eukaryotes shown in Figure 5.7. However, virtually none of the annotations can be fully covered, i.e., it is not possible to solve structures of all clusters for a given annotation. This reveals that nearly all annotations include proteins with difficult to solve structures. Panel B in Figure 5.8 shows that current coverage of functional annotations based on the structures of human proteins from the PDB is fairly low, e.g., 10% of annotations are covered at 35%, and 90% at 6%. However, using X-ray crystallography and homology modeling, this coverage could be substantially improved in the future (based on the median-based cutoff on the crystallization propensity scores) to 90% of annotations at 62% coverage, and 50% at 77%.

5.4.6 Analysis of GPCRs

GPCRs were clustered at 30% identity and results were processed per-cluster; we selected one representative sequence with the highest crystallization propensity score from each cluster.

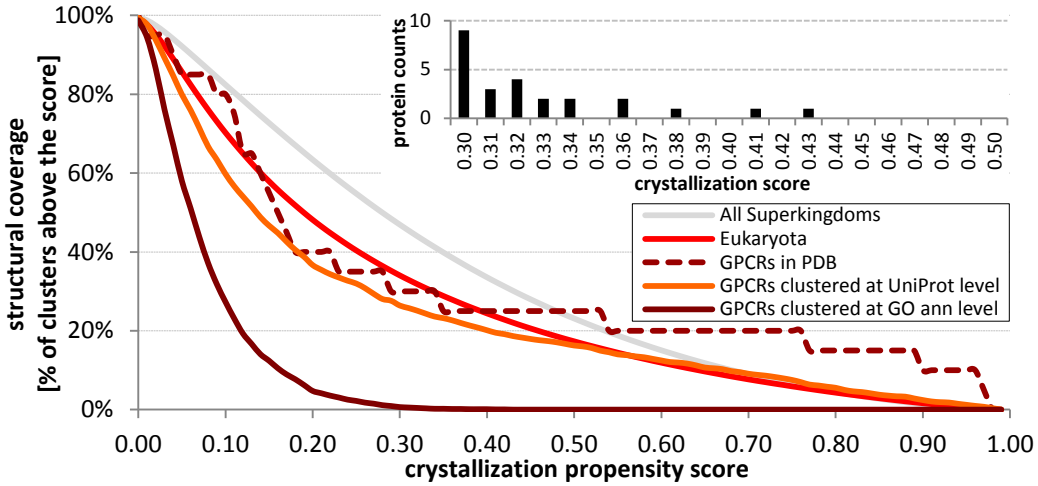


Figure 5.9: Analysis of crystallization propensities of G-protein Coupled Receptors.
Y-axis shows the % of annotations which have at least x% of clusters covered, where x is displayed on x-axis. GPCRs in PDB show the distribution of crystallization propensity scores for GPCRs in the PDB. The inset shows distribution of GPCRs with scores above 0.3.

Figure 5.9 shows the crystallization propensity scores for the GPCRs. These scores are substantially lower than for a generic set of eukaryotic chains. GPCRs that are already solved (in the PDB) have a relatively high predicted crystallization propensity,

which shows that they could be potentially found by fDETECT and that so far relatively easier targets were solved. However, some of these solved GPCRs are engineered protein fragments, which results in higher crystallization scores.

The use of X-ray crystallography and homology modeling (orange line/GPCRs clustered using chains in UniProt) would significantly improve structural coverage of GPCRs. Moreover (see inset in Figure 5.9), we found a number of GPCR chains with high crystallization propensity scores for which structures should be easier to solve, i.e., 26 out of the considered 24,730 GPCRs have fairly high scores at over 0.3. This list is available in Appendix E.

5.4.7 Summary of the analysis

We observed that different organisms differ in their difficulty for structural determination and these differences can be aggregated at the superkingdom level; archaea proteomes are overall the easiest to solve, although their crystallization propensity scores overlap with the scores of some bacterias and bacterial/archaea viruses. Eukaryotes and eukaryotic viruses are the hardest to solve and their scores are lower and have little overlap with the other superkingdoms (see **Characterization of crystallization propensity for proteomes**). Similar observations are true when we allow homology modeling across organisms, with the differences that values of coverage are proportionally higher and the fact that archaea and bacterial organisms overlap to a greater extend. Moreover, use of more advanced, future homology modeling methods that could solve structures at lower sequence identity threshold would still lead to higher coverage for bacterial and archea organisms and viruses that they host, compared to substantially lower coverage for eukaryotic organisms and their viruses. The current structural coverage (based on all solved structures in PDB) also results in a similar distribution of coverage values between the superkingdoms, with the values being proportionally smaller. The main difference compared to the results for the attainable coverage is a bigger overlap between values for the archaea, bacterial and eukaryotic organisms. However, these differences may not solely reflect crystallization propensity, but they are possibly also influenced by research interests in individual superkingdoms (see **Attainable structural coverage**).

Using a cut-off score at median of the crystallization scores obtained for solved chains from PDB, we found that using current crystallization protocols we can obtain structures for about 43% of all considered protein clusters (and each cluster can be covered using homology modeling from one of its members), with the average structural coverage rates that vary substantially between superkingdoms: 33% for eukaryotes and eukaryotic viruses, 58% for bacterial organisms, and 68% for archaea proteomes and archeal/bacteria viruses (see **Attainable structural coverage of protein families**). Moreover, majority of bacterial and archaea organisms and bacterial/archaea viruses can be covered at >70%, eukaryotic organisms could be covered at substantially lower rates, between 25 and around 70%, while eukaryotic viruses have the lowest attainable coverages which ranges between 10 and 40% (see **Attainable structural coverage**).

We also show that use of target prediction methods, such as our fDETECT, can significantly increase the attainable coverage when compared with a "random" approach that does not utilize crystallization scores. The increase in the structural coverage when using target selection ranges between 25 and 40%, which depends on the coverage values of the "random" approach, i.e., the lower the "random" coverage the higher the increase obtained by utilization of the target selection. However, for some eukaryotic viruses this improvement is smaller, between 5 and 25%. As expected, further progress in homology modeling, expressed by ability to model structures of proteins with lower levels of sequence identity, would result in substantial increase in the attainable structural coverage, pushing the coverage for almost all proteomes to over 90% when assuming modeling would be successful at 20% identity; an exception are eukaryotic viruses whose coverage would be in most cases in the 80 to 90% range (see **Attainable structural coverage**).

The analysis of the structural coverage of functional and localization-based annotations defined in GO reveals that we can currently solve at least one structure for each annotation, The fraction of annotations for which at least half of the clusters can be solved varies widely between superkingdoms, with low 6% in viruses, and higher values in the other superkingdoms reaching 79% in eukaryotes, 88% in bacteria and 95% for archaea organisms (see **Attainable structural coverage of**). Moreover, the number of annotations that could be fully structurally covered (those for which all clusters can

be solved) is low, between 0% for eukaryotes and viruses, and 3 to 8% for bacteria and archaea, respectively. This low coverage means that almost all annotations contain some hard to crystallize proteins, and points out the necessity to develop new strategies for structure determination.

The inspection of the attainable coverage for *Homo sapiens* proteome shows that it is the proteome with the highest crystallization propensity among eukaryotes. The reason for that may be that since this proteome attracts more attention and resources being of the primary importance for us. We showed that, utilizing the available structures and homology modeling, we now have structures for about 14% of clusters of human proteins (which corresponds to 26% of individual proteins), and using current X-ray crystallography techniques and homology modeling we should be able to obtain structures for around 70% of the clusters from the human proteome. We can further increase this coverage to around 93% given that the homology modeling could be improved to model structures at 20% sequence identity (see **Attainable structural coverage**). Analysis of the functional annotations in human proteins shows trends that are similar to the other proteomes, i.e., we can obtain at least one structure for all annotations, solve over a half of the proteins for over 95% of annotations, and fully structurally cover only around 1% of the annotations. We also show that the coverage of GO annotations can be greatly improved using current crystallization and homology modeling technologies (see **Analysis of human proteome**).

Finally, we analyzed peculiarities of the structural coverage for an interesting family of G protein coupled receptors (GPCRs), which are found primarily in eukaryotes. Our study demonstrates that their crystallization propensity is relatively low, with around 72% of these proteins having the crystallization score below 0.1; to compare, less than 20% of eukarotic proteins have such low scores. Nevertheless, we fund that use of homology modeling could substantially increase the structural coverage of this protein family. We also investigated the crystallization propensities of GPCRs for which structures were deposited to PDB and found that, as expected, their scores are higher compared to overall scores for all GPCRs. This means that so far easier GPCR targets were solved, but this result is also influenced by the fact that some of these proteins were engineered to enhance their crystallization propensity. We provided a list of 26

GPCR targets with the crystallization propensity scores of at least 0.3, which we believe should be easier to solve (see **Analysis of GPCRs**).

5.5 Conclusions

We designed a novel, accurate and time-efficient crystallization propensity predictor fDETECT to perform a large-scale analysis of the attainable structural coverage. Our predictor uses features and prediction model which can be quickly calculated from the input protein sequence. As a result of that, fDETECT is six orders of magnitude faster than the currently best performing method while achieving a competitive predictive performance. To further demonstrate predictive power of our method, we studied predicted crystallization propensity scores for a large set of over a 50,000 non-redundant chains from PDB. This empirical analysis confirmed that our method on average predicts proteins that are already crystallized as easier than those which do not yet have crystal structures. Interestingly, crystallization propensities predicted by fDETECT correlate with resolutions of the resulting crystals, which suggests that our method could be used to find proteins for which high-resolution crystals can be generated.

fDETECT uses only eleven features. These features were carefully selected from among 1,283 features that were found to be correlated with crystallization propensity and which were characterized by relatively low cross-correlation. The strong predictive performance of our method suggests that relatively basic and fast to compute proteins characteristics are sufficient for this type of prediction. Our selected features, except for one, are connected with protein characteristics which were used in previous studies and were found to be useful for crystallization propensity predictions; fDETECT combines them together in a way that optimizes predictive performance. The one novel feature is possibly associated with interference of the protein chain with N-terminus affinity tags, but a more detailed investigation is needed to substantiate this claim. Similarly, a further research is needed to validate our model-driven hypothesis that the non-crystallizable targets have segments of a broader range of free energy than the crystallizable targets.

Utilizing computational efficiency of fDETECT we analyzed crystallization propensity scores and the resulting structural coverage for close to two thousands proteomes, and all functional and localization annotations available in Gene Ontology for these proteomes. We clustered our dataset at 30% sequence identity threshold to model the use of homology modeling. Our analysis revealed that crystallization propensities vary between organisms and while analyzing organisms' median crystallization scores we found a bimodal distribution where archaeal and bacterial proteomes are the easier to crystallize and eukaryotic proteomes are harder. The crystallization propensity of viral proteins depends on viruses' host, i.e., these proteins are easier to solve for viruses with archaeal and bacterial hosts, and harder for viruses with eukaryotic hosts. Our study show that current X-ray crystallography combined with homology modeling could provide an average, over all considered organisms, structural coverage of 73% with over 65% for archaea and bacteria, over 50% for archaeal and bacterial viruses, between 25 and 70% for eukaryotes, and below 35% for eukaryotic viruses. At least one structure can be determined for each GO annotation. However, 7, 80, 90 and 95% of these functional and localization annotations could attain 50% structural coverage in viruses, eukaryotes, bacteria and archaea, respectively. The structural coverage could be substantially increased to an average 80% and 96% given that homology modeling would be successful at 25% and 20% sequence identity, respectively. Moreover, use of knowledge-based target selection increases coverage by a significant margin which for majority of the organisms is between 25 to 40%, when compared to an approach where proteins are chosen at random. We also showed that human proteome is one of eukaryotes with highest attainable structural coverage and using current techniques its coverage could reach around 70%; this can be further improved to up to 93% if homology modeling would be successful at the 20% sequence identity. Finally, we show that GPCRs are hard to crystallize, as over 72% of these proteins have crystallization score below 0.1, but we were still able to find a couple dozen GPCR targets with promising scores of over 0.3.

Our analysis should be considered in the context of several assumptions. First, the dataset used to develop and evaluate our method contains individual proteins taken directly from TargetDB with the removed affinity tags. This means that fDETECT is not

capable of predicting crystallization propensity of protein complexes. It uses a single protein chain as the input, which implies that it does not consider cofactors such as ligands. Moreover, we utilize the sequences in their wild form, which means that we do not consider modifications (mutations, affinity tags, etc.) that could enhance crystallization propensity. Finally, homology modeling is assumed to produce structures with sufficient quality at 30% sequence identity which may not be true in all cases.

Chapter 6

Summary and conclusions

Our main aim is prediction and characterization of protein crystallization propensity from AA sequence. This is an important problem for which solutions are sought by structural genomics centers to increase structure determination rate and lower the associated costs of protein structure determination. We started our journey in this field of research by designing CRYSTALP2 classifier, which at the time of publication offered prediction quality that was comparable to existing predictors. Next, we analyzed predictive performance of the existing methods and we drew conclusions that some of these methods are complementary to each other. We took advantage of this finding and designed MetaPPCP, a meta-predictor which uses outputs of other methods as well as some of their inputs, to predict crystallization propensity scores. When empirically evaluated, this method turned up to have the highest predictive performance. After developing MetaPPCP, we executed a study where we evaluated value of various protein characteristics and predicted structural information for the prediction of crystallization outcomes. To investigate novel protein characteristics we utilized AAIndex database which stores over 500 AA indices describing various physiochemical properties, whereas the structural information was represented by predicted relative solvent accessibility and intrinsic disorder. We used these data sources to design features that we then used to develop a new crystallization propensity predictor CRYSpred. The method had predictive quality similar to MetaPPCP, and used 15 features/characteristics that were well aligned with existing observations. Our predictor was arguably the first to combine these features together to offer strong predictive performance.

While analyzing results of various crystallization propensity predictors on our test datasets we found out that the results on the datasets with newer targets were substantially lower than for the datasets with older targets. We also found that these

crystallization propensity predictions give limited information to crystallographers in case of negative (i.e., protein is predicted to fail crystallization) predictions. To this end, we developed a new method that not only improves accuracy of the predictions (compared with other relevant predictors) but also gives additional information in the case of the negative outcome. First, we designed a novel protocol to obtain and annotate crystallization trials from PepcDB to define outcomes that expand the information about negative predictions. Beside the positive class label (crystallization successful), these annotations include three negative labels which correspond to failures at each of the three main crystallization steps: production of protein material, purification, and crystallization. We then designed a first-of-its-kind crystallization propensity predictor, PPCpred which provides more feedback to crystallographers by predicting the step in the crystallization pipeline that is the most likely to be the cause of the failure. Empirical analysis of the new predictor demonstrated that its predictive performance is better than the performance of existing predictors. It also showed that prediction performance of existing predictors deteriorated over time, which supported our earlier hypothesis. However, PPCpred utilized some input features based on protein's predicted structural information, which are relatively time consuming to compute. Consequently, our predictor could not be applied to predict large protein datasets.

The exponential increase in the number of known protein sequences and lack of a large-scale study that would analyze propensity of crystallization for a comprehensive set of proteins motivated us to the development of new time efficient method, called fDETECT. This method turned out to have predictive performance that was empirically shown to be comparable to PPCpred, while being six orders of magnitude faster. This high time efficiency allowed us to analyze crystallization propensity for close to ten million proteins from almost 2,000 fully sequenced proteomes available in the UniProt databank. This analysis was the first to estimate the attainable structural coverage of various proteomes, superkingdoms and functional annotations which can be obtained by combining current X-ray protocols and homology modeling.

6.1 Major contributions

The major contributions of this research include:

- Goal 1. To develop methods which provide more accurate prediction of crystallization propensity when compared to existing predictors.
    - participation in the design and empirical evaluation of a novel crystallization propensity predictor CRYSTALP2, creation of new test dataset
    - development of a publicly available webserver for CRYSTALP2
    - empirical analysis of complementarity of four crystallization propensity predictors: OB-Score, ParCrys, XtalPred, and CRYSTALP2
    - development and empirical assessment of novel meta predictor MetaPPCP
    - development and empirical assessment of novel crystallization propensity predictor CRYSpred
    - analysis of the protein characteristics related to protein crystallization based on the features used by CRYSpred
- Goal 2. To predict outcomes of individual steps in the crystallization protocol.
    - design of a first-of-its kind protocol to obtain and annotate crystallization trials from PepcDB
    - development and empirical assessment of a first-of-its kind method PPCpred, which predicts outcomes for four steps in the crystallization pipeline
    - comprehensive empirical evaluation of predictive performance of PPCpred and existing crystallization propensity predictors, which included analysis of the predictive performance over time
    - analysis of protein characteristics related to crystallization based on features that are used by PPCpred
    - development of a publicly available webserver for PPCpred

- Goal 3. To compute and analyze of the attainable structural coverage.
    - development of time-efficient and accurate crystallization propensity predictor, fDETECT
    - comprehensive empirical evaluation of fDETECT which includes comparison with existing predictors and analysis of the method's predictions on large non-redundant set of crystallized proteins from PDB.
    - analysis of the protein characteristics related to protein crystallization based on the features used by fDETECT
    - first-of-its kind large scale and exhaustive analysis of an attainable structural coverage of a current snapshot of protein universe, using X-ray crystallography and homology modeling, which includes analysis at the proteome and functional annotation levels.
    - study of crystallization propensities of G protein-coupled receptors and application of fDETECT to provide the most suitable targets from this protein family for structure determination

The work presented in this thesis was published in the following publications:

- Kurgan, L.A., Razib, A.A., Aghakhani, S., Dick, S., Mizianty, M.J. & Jahandideh, S. (2009). CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC structural biology* 9 p. 50.
- Kurgan, L.A. & Mizianty, M.J. (2009). Sequence-Based Protein Crystallization Propensity Prediction for Structural Genomics: Review and Comparative Analysis. *Natural Science* 1 (2) pp. 93–106.
- Mizianty, M.J. & Kurgan, L.A. (2009). Meta prediction of protein crystallization propensity. *Biochemical and biophysical research communications* 390 (1) pp. 10–15.
- Mizianty, M.J. & Kurgan, L.A. (2011). Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics* 27 (13) pp. i24–i33.

- Mizianty, M.J. & Kurgan, L.A. (2012). CRYSpred: Accurate Sequence-Based Protein Crystallization Propensity Prediction Using Sequence-Derived Structural Characteristics. *Protein and peptide letters* 19 (1) pp. 40–49.

6.2 Major findings

By analyzing predictive performance of our five predictors and other existing methods we showed that their predictive performance deteriorates over time, which most likely is associated with the fact that crystallization protocols are being improved to crystallize previously non-crystallizable targets. We also discovered that it is possible to accurately predict outcomes of protein material production and protein purification; these are major steps in the protein crystallization pipelines. Moreover, during design of our newest fDETECT method, we demonstrated that it is possible to build accurate method which is also runtime efficient, and we concluded that although features based on predicted structural information are correlated with crystallization output they may be replaced by features which can be quickly calculated from protein sequence.

We confirmed a several protein characteristics that are linked to protein crystallization propensity. These include hydrophobicity, charge, pI, protein intrinsic disorder, relative solvent accessibility, and content of certain AAs (Arg, Asn, Cys, Glu, His, Met, and Ser). We also found that some of these features in combination with additional details (e.g., related to the solvent accessibility of selected residues types) provide useful markers of crystallization. Our predictive models implement a novel way of combining these factors to obtain good predictive performance. We also found out that protein properties based on AA indices related to free energy can be used to predict crystallization outcomes, where non-crystallizable proteins potentially have segments of broader range of free energies, and that position of AAs of medium polarizability close to N-terminus possibly hinders crystallization. The latter effect may be connected with some interactions with affinity tags that are often placed at the N-terminus. We plan further investigations of these interesting findings.

Our large scale analysis of predicted crystallization propensity for the current snapshot of the protein universe demonstrated that organisms are divided into two groups: easy to crystallize that includes all archaean and bacterial proteomes and

archaeal/bacterial viruses, and hard to crystallize which include eukaryotes and eukaryotic viruses. We also showed that current X-ray crystallography combined with homology modeling could provide an average structural coverage of 73% with over 60% coverage for archaea and bacteria, and between 35 and 70% for eukaryotes. Moreover, we demonstrated that use of knowledge-based target selection increases coverage by a significant margin, which for majority of the considered proteomes is between 25 and 40%. Finally, our analysis revealed that human proteome has one of the highest attainable structural coverage values among eukaryotic proteomes and using current X-ray crystallography protocols and homology modeling it can attain structural coverage of 70%.

6.3 Future work

As we found in this thesis, crystallization propensity predictors need continuing improvements to keep up with the advances in crystallization protocols. Besides the obvious improvements, like those based on the use of new data, application of new machine learning algorithms, and designing new input features, we also define three specific directions that we may consider to investigate in the future. They include suggestion of crystallization enabling mutations, prediction of protein solubility, and suggestion of the most feasible crystallization protocols based on propensity predictions tailored to specific crystallization protocols.

6.3.1 Crystallization enabling mutations

Crystallization enabling mutations are mutations in protein sequence which make them more likely to crystallize. One of such protein engineering approaches, called Surface Entropy Reduction (SER), was proposed in 2004 by prof. Zygmunt Derewenda. In this method clusters of two or three AAs with high conformational entropy are replaced by AAs with a lower entropy (Derewenda, 2004). This study leaded to the development of the SER server (Goldschmidt et al., 2007). This server proposes AAs clusters mutations to make crystallization easier. Some successful applications of the mutations proposed by SER have been reported, making this server a valuable tool in structural genomics. However, this approach uses only SER to find the potential mutations. Our approach would make use of TargetTrack database to develop a database of mutations that we could use to build a new predictor. To create the database, we would select all pairs of

similar proteins where one protein is crystallizable, and the second is not. We could then annotate differences in the protein pairs and use them to train machine learning algorithm to recognize potential mutations that enhance crystallization propensity. The set of mutations may be further expanded by selecting pairs of crystallizable protein from PDB, which have crystal resolutions that differ by a large margin. Such tool would possibly improve structure determination rate, and could also provide new valuable insights into protein characteristics associated with crystallization.

6.3.2 Prediction of solubility propensity

Beside crystallization, another substantial bottleneck in the determination of protein structure is expression of soluble proteins. It is estimated that around 50% of proteins expressed in *E. coli.* are insoluble (Kim et al., 2011). This step is partially addressed in our prediction of the three steps of the crystallization pipeline; however, solubility does not have a distinctive stop status in PepcDB and TargetTrack, and for some samples it is impossible to determine whether they are soluble or insoluble. Protein solubility has been addressed earlier than the studies of protein crystallization propensities; the first approach for solubility prediction was proposed over 20 years ago (Wilkinson & Harrison, 1991). Since then the amount of research was relatively limited. Only a handful of methods were developed although interest in this prediction was fueled by large scale SG experiments that were carried around eight years ago (Davis et al., 1999; Bertone et al., 2001; Goh et al., 2004; Idicula-Thomas & Balaji, 2005; Idicula-Thomas et al., 2006; Smialowski et al., 2007; Magnan et al., 2009; Agostini et al., 2012; Smialowski et al., 2012). Solubility step roughly corresponds to the protein material production step which is predicted by PPCpred, as it includes most of the proteins that failed do dissolve. However, some of such proteins, due to limitations of the PepcDB annotations, may right now be labeled as failed to purify. In our future work solubility may replace the production of protein material step that we currently utilize.

6.3.3 Protocol suggestion system

Along abovementioned improvements, we believe that major advances may be achieved by designing solubility, purification, and crystallization predictors that are tailored to specific crystallization protocols. Right now some protocols may dissolve/purify/crystallize proteins which would fail in others protocols as they are using

different procedures and conditions. In most SG centers standardized and cost effective approaches are tried first and in case of failure some more sophisticated approaches may be applied or a target is abandoned. Designing predictors that would be able to predict propensities for solubility, purification or crystallization for different protocols, and then suggesting the protocol which will be the most likely to successfully complete a given task might further increase structure determination success rates as the targets would be addressed directly by appropriate protocols. This goal may be achieved by creating multiple protein sets, divided with respect to the protocol used, and creating a predictor for each of them. Some information about the used protocols is already available in the TargetTrack database and could be used to develop these protein sets. Beside per set (protocol) evaluation of the corresponding predictors, an evaluation on the interception of these sets, i.e., for the sets of proteins tried by all considered protocols, would be necessary to estimate how accurately these protocols are suggested. Such analysis may be problematic however, as there may be not enough proteins which were tried by multiple protocols.

Bibliography

Agostini, F., Vendruscolo, M. & Tartaglia, G.G. (2012). Sequence-Based Prediction of Protein Solubility. *Journal of Molecular Biology* 421 pp. 237–241

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25 (17) pp. 3389–3402

Anfinsen, C.B. (1973). Principles that Govern the Folding of Protein Chains. *Science* 181 (4096) pp. 223–230

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25 (1) pp. 25–29

Babnigg, G. & Joachimiak, A. (2010). Predicting protein crystallization propensity from protein sequence. *Journal of Structural and Functional Genomics* 11 (1) pp. 71–80

Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science (New York, N.Y.)* 294 (5540) pp. 93–96

Beadle, W.G. & Tatu, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences* 27 (11) pp. 499–506

Berman, H.M., Kleywegt, G.J., Nakamura, H. & Markley, J.L. (2012). The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure (London, England : 1993)* 20 (3) pp. 391–396

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Research* 28 (1) pp. 235–242

Bernal, J. & Crowfoot, D. (1934). X-ray photographs of crystalline pepsin. *Nature* 133 pp. 794 – 795

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T. & Gerstein, M. (2001). SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic acids research* 29 (13) pp. 2884–2898

Biertumpfel, C., Basquin, J. & Suck, D. (2005). Practical implementations for improving the throughput in a manual crystallization setup. *Journal of Applied Crystallography* 38 (3) pp. 568–570

Black, S.D. & Mould, D.R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Analytical biochemistry* 193 (1) pp. 72–82

Brenner, S.E. (2000). Target selection for structural genomics. *Nature structural biology* 7 Suppl pp. 967–969

Brenner, S.E., Barken, D. & Levitt, M. (1999). The PRESAGE database for structural genomics. *Nucleic acids research* 27 (1) pp. 251–253

Bugmann, G. (1998). Normalized Gaussian Radial Basis Function networks. *Neurocomputing* 20 (1-3) pp. 97–110

Bull, H.B. & Breese, K. (1974). Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Archives of Biochemistry and Biophysics* 161 (2) pp. 665–670

Canaves, J.M., Page, R., Wilson, I.A. & Stevens, R.C. (2004). Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics. *Journal of molecular biology* 344 (4) pp. 977–991

Casari, G. & Sippl, M.J. (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *Journal of molecular biology* 224 (3) pp. 725–732

Cessie, L. & van Houwelingen, J. (1992). Ridge Estimators in Logistic Regression. *Applied Statistics* 41(1) pp. 191 – 201

Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.-S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M. & Wang, L.K. (2002). Structural genomics: a pipeline for providing structures for the biologist. *Protein science : a publication of the Protein Society* 11 (4) pp. 723–738

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

Chandonia, J.-M., Kim, S.-H. & Brenner, S.E. (2006). Target selection and deselection at the Berkeley Structural Genomics Center. *Proteins* 62 (2) pp. 356–370

Chang, C.-C. & Lin, C.-J. (2011). {LIBSVM}: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (3) p. 27

Chang, R.L., Andrews, K., Kim, D., Li, Z., Godzik, A. & Palsson, B.O. (2013). Structural Systems Biology Evaluation of Metabolic Thermotolerance in Escherichia coli. *Science* 340 (6137) pp. 1220–1223

Chayen, N.E. (2004). Turning protein crystallisation from an art into a science. *Current opinion in structural biology* 14 (5) pp. 577–583

Chen, K. & Kurgan, L.A. (2009). Investigation of atomic level patterns in protein--small ligand interactions. *PloS one* 4 (2) p. e4473

Chen, K., Kurgan, L.A. & Rahbari, M. (2007). Prediction of protein crystallization using collocation of amino acid pairs. *Biochemical and biophysical research communications* 355 (3) pp. 764–769

Chen, K., Mizianty, M.J., Gao, J. & Kurgan, L.A. (2011). A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure London England 1993* 19 (5) pp. 613–621

Chen, L., Oughtred, R., Berman, H.M. & Westbrook, J. (2004). TargetDB: a target registration database for structural genomics projects. *Bioinformatics* 20 (16) pp. 2860–2862

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M. & Arrowsmith, C.H. (2000). Structural proteomics of an archaeon. *Nature structural biology* 7 (10) pp. 903–909

Creamer, T.P. (2000). Side-chain conformational entropy in protein unfolded states. *Proteins* 40 (3) pp. 443–450

Davis, G.D., Elisee, C., Newham, D.M. & Harrison, R.G. (1999). New Fusion Protein Systems Designed to Give Soluble Expression in Escherichia coli. *Biotechnology and Bioengineering* 65 (4) pp. 382–388

Derewenda, Z.S. (2004). Rational protein crystallization by mutational surface engineering. *Structure (London, England : 1993)* 12 (4) pp. 529–535

Dor, O. & Zhou, Y. (2007). Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins* 66 (4) pp. 838–845

Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21 (16) pp. 3433–3434

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I. & Kim, S.H. (1999). Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins* 35 (4) pp. 401–407

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)* 26 (19) pp. 2460–2461

Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* 319 (6050) pp. 199–203

Ellis, J.J., Broom, M. & Jones, S. (2007). Protein-RNA interactions: structural analysis and functional classes. *Proteins* 66 (4) pp. 903–911

Faraggi, E., Xue, B. & Zhou, Y. (2009). Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins* 74 (4) pp. 847–856

Fernàndez-Busquets, X., De Groot, N.S., Fernandez, D. & Ventura, S. (2008). Recent structural and computational insights into conformational diseases. *Current Medicinal Chemistry* 15 (13) pp. 1336–1349

Fogg, M.J. & Wilkinson, A.J. (2008). Higher-throughput approaches to crystallization and crystal structure determination. *Biochemical Society transactions* 36 (Pt 4) pp. 771–775

Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W. & Bryant, S.H. (2010). The NCBI BioSystems database. *Nucleic acids research* 38 (Database issue) pp. D492–6

Geerlof, A., Brown, J., Coutard, B., Egloff, M.P., Enguita, F.J., Fogg, M.J., Gilbert, R.J.C., Groves, M.R., Haouz, A., Nettleship, J.E., Nordlund, P., Owens, R.J., Ruff, M., Sainsbury, S., Svergun, D.I. & Wilmanns, M. (2006). The impact of protein characterization in structural proteomics. *Acta crystallographica. Section D, Biological crystallography* 62 (Pt 10) pp. 1125–1136

Gerdts, C.J., Tereshko, V., Yadav, M.K., Dementieva, I., Collart, F., Joachimiak, A., Stevens, R.C., Kuhn, P., Kossiakoff, A. & Ismagilov, R.F. (2006). Time-controlled microfluidic seeding in nL-volume droplets to separate nucleation and growth stages of protein crystallization. *Angewandte Chemie (International ed. in English)* 45 (48) pp. 8156–8160

Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Current opinion in structural biology* 16 (2) pp. 172–177

Goh, C.-S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H. & Gerstein, M. (2004). Mining the structural genomics pipeline:

identification of protein properties that affect high-throughput experimental analysis. *Journal of molecular biology* 336 (1) pp. 115–130

Goh, C.-S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.-C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G.T. & Gerstein, M. (2003). SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic acids research* 31 (11) pp. 2833–2838

Goldsack, D.E. & Chalifoux, R.C. (1973). Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. *Journal of theoretical biology* 39 (3) pp. 645–651

Goldschmidt, L., Cooper, D.R., Derewenda, Z.S. & Eisenberg, D. (2007). Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein science : a publication of the Protein Society* 16 (8) pp. 1569–1576

Grey, J. & Thompson, D. (2010). Challenges and Opportunities for New Protein Crystallization Strategies in Structure-Based Drug Design. *Expert opinion on drug discovery* 5 (11) pp. 1039–1045

Gront, D., Grabowski, M., Zimmerman, M.D., Raynor, J., Tkaczuk, K.L. & Minor, W. (2012). Assessing the accuracy of template-based structure prediction metaservers by comparison with structural genomics structures. *Journal of structural and functional genomics* 13 (4) pp. 213–225

Guido, R.V.C., Oliva, G. & Andricopulo, A.D. (2008). Virtual screening and its integration with modern drug design technologies. *Current Medicinal Chemistry* 15 (1) pp. 37–46

Guruprasad, K., Reddy, B. V & Pandit, M.W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein engineering* 4 (2) pp. 155–161

Hall, M.A. (1999). *Correlation-based Feature Selection for Machine Learning*. The University of Waikato.

Hall, M.A., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter* 11 (1) p. 10

Harrison, S.C. (2004). Whither structural biology? *Nature structural & molecular biology* 11 (1) pp. 12–15

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica* 47 (1) pp. 153–161

Hopp, T.P. & Woods, K.R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences of the United States of America* 78 (6) pp. 3824–3828

Hui, R. & Edwards, A.M. (2003). High-throughput protein crystallization. *Journal of structural biology* 142 (1) pp. 154–161

Idicula-Thomas, S. & Balaji, P. V (2005). Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein science* 14 (3) pp. 582–592

Idicula-Thomas, S., Kulkarni, A.J., Kulkarni, B.D., Jayaraman, V.K. & Balaji, P. V (2006). A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics* 22 (3) pp. 278–284

Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *Journal of biochemistry* 88 (6) pp. 1895–1898

Jahandideh, S. & Mahdavi, A. (2012). RFCRYS: sequence-based protein crystallization propensity prediction by means of random forest. *Journal of theoretical biology* 306 pp. 115–119

Joachimiak, A. (2009). High-throughput crystallography for structural genomics. *Current opinion in structural biology* 19 (5) pp. 573–584

John, G. & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In: *In proceedings of the eleventh conference on uncertainty in artificial intelligence*. 1995, San Mateo, CA, USA: Morgan Kaufmann Publishers Inc., pp. 338 – 345.

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

Juretić, D., Lučić, B., Zucić, D. & Trinajstić, N. (1998). *Theoretical Organic Chemistry* Elsevier.

Kandaswamy, K.K., Pugalenthi, G., Suganthan, P.N. & Gangal, R. (2010). SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. *Protein and Peptide Letters* 17 (4) pp. 423–430

Kantardjieff, K.A., Jamshidian, M. & Rupp, B. (2004). Distributions of pI versus pH provide prior information for the design of crystallization screening experiments: response to comment on “Protein isoelectric point as a predictor for increased crystallization screening efficiency.” *Bioinformatics* 20 (14) pp. 2171–2174

Kantardjieff, K.A. & Rupp, B. (2004). Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics* 20 (14) pp. 2162–2168

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. & Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic acids research* 36 (Database issue) pp. D202–205

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. & Phillips, D.C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* 181 (4610) pp. 662–666

Kidera, A., Konishi, Y., Oka, M., Ooi, T. & Scheraga, H.A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry* 4 (1) pp. 23–55

Kim, Y., Babnigg, G., Jedrzejczak, R., Eschenfeldt, W.H., Li, H., Maltseva, N., Hatzos-Skintges, C., Gu, M., Makowska-Grzyska, M., Wu, R., An, H., Chhor, G. & Joachimiak, A. (2011). High-throughput protein purification and quality assessment for crystallization. *Methods* 55 (1) pp. 12–28

Kim, Y., Dementieva, I., Zhou, M., Wu, R., Lezondra, L., Quartey, P., Joachimiak, G., Korolev, O., Li, H. & Joachimiak, A. (2004). Automation of protein purification for structural genomics. *Journal of structural and functional genomics* 5 (1-2) pp. 111–118

Klebe, G. (2000). Recent developments in structure-based drug design. *Journal of molecular medicine (Berlin, Germany)* 78 (5) pp. 269–281

Koonin, E. V, Wolf, Y.I. & Karev, G.P. (2002). The structure of the protein universe and genome evolution. *Nature* 420 (6912) pp. 218–223

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E. & Berman, H.M. (2006). The RCSB PDB information portal for structural genomics. *Nucleic acids research* 34 (Database issue) pp. D302–305

Kurgan, L.A. & Mizianty, M.J. (2009). Sequence-Based Protein Crystallization Propensity Prediction for Structural Genomics: Review and Comparative Analysis. *Natural Science* 01 (02) pp. 93–106

Kurgan, L.A., Razib, A.A., Aghakhani, S., Dick, S., Mizianty, M.J. & Jahandideh, S. (2009). CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC structural biology* 9 p. 50

Landwehr, N., Hall, M.A. & Frank, E. (2005). Logistic Model Trees. *Machine Learning* 59 (1-2) pp. 161–205

Lawson, E.Q., Sadler, A.J., Harmatz, D., Brandau, D.T., Micanovic, R., MacElroy, R.D. & Middaugh, C.R. (1984). A simple experimental model for hydrophobic interactions in proteins. *The Journal of biological chemistry* 259 (5) pp. 2910–2912

Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L.S., Miller, M.D., McPhillips, T.M., Miller, M.A., Scheibe, D., Canaves, J.M., Guda, C., Jaroszewski, L., Selby, T.L., Elsliger, M.-A., Wooley, J., Taylor, S.S., Hodgson, K.O., Wilson, I.A., Schultz, P.G. & Stevens, R.C. (2002). Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. *Proceedings of the National Academy of Sciences of the United States of America* 99 (18) pp. 11664–11669

Levitt, M. (2007). Growth of novel protein structural data. *Proceedings of the National Academy of Sciences of the United States of America* 104 (9) pp. 3183–3188

Liu, J. & Rost, B. (2002). Target space for structural genomics revisited. *Bioinformatics (Oxford, England)* 18 (7) pp. 922–933

Livingstone, C.D. & Barton, G.J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer applications in the biosciences : CABIOS* 9 (6) pp. 745–756

Lundstrom, K. (2009). An overview on GPCRs and drug discovery: structure-based drug design and structural biology on GPCRs. *Methods in molecular biology (Clifton, N.J.)* 552 pp. 51–66

Luscombe, N.M., Laskowski, R.A. & Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic acids research* 29 (13) pp. 2860–2874

Maeki, M., Yoshizuka, S. & Yamaguchi, H. (2012). X-ray Diffraction of Protein Crystal Grown in a Nano-liter Scale Droplet in a Microchannel and Evaluation of Its Applicability. *Analytical Sciences* 28 (January) pp. 65–68

Magnan, C.N., Randall, A. & Baldi, P. (2009). SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* 25 (17) pp. 2200–2207

Manavalan, P. & Ponnuswamy, P.K. (1978). Hydrophobic character of amino acid residues in globular proteins. *Nature* 275 (5681) pp. 673–674

Mariani, V., Kiefer, F., Schmidt, T., Haas, J. & Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins* 79 Suppl 1 pp. 37–58

Mizianty, M.J. & Kurgan, L.A. (2012). CRYSpred: Accurate Sequence-Based Protein Crystallization Propensity Prediction Using Sequence-Derived Structural Characteristics. *Protein and peptide letters* 19 (1) pp. 40–49

Mizianty, M.J. & Kurgan, L.A. (2009). Meta prediction of protein crystallization propensity. *Biochemical and biophysical research communications* 390 (1) pp. 10–15

Mizianty, M.J. & Kurgan, L.A. (2011). Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics* 27 (13) pp. i24–i33

Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)--round IX. *Proteins* 79 Suppl 1 pp. 1–5

Nair, R., Liu, J., Soong, T.-T., Acton, T.B., Everett, J.K., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., Montelione, G.T. & Rost, B. (2009). Structural genomics is the largest contributor of novel structural leverage. *Journal of structural and functional genomics* 10 (2) pp. 181–191

Norin, M. & Sundström, M. (2001). Protein models in drug discovery. *Current opinion in drug discovery & development* 4 (3) pp. 284–290

Oldfield, C.J., Ulrich, E.L., Cheng, Y., Dunker, A.K. & Markley, J.L. (2005). Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins* 59 (3) pp. 444–453

Oobatake, M. & Ooi, T. (1977). An analysis of non-bonded energy of proteins. *Journal of Theoretical Biology* 67 (3) pp. 567–584

Overton, I.M. & Barton, G.J. (2006). A normalised scale for structural genomics target ranking: the OB-Score. *FEBS letters* 580 (16) pp. 4005–4009

Overton, I.M., van Niekerk, C.A.J. & Barton, G.J. (2011). XANNpred: neural nets that predict the propensity of a protein to yield diffraction-quality crystals. *Proteins* 79 (4) pp. 1027–1033

Overton, I.M., Padovani, G., Girolami, M. a & Barton, G.J. (2008). ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics* 24 (7) pp. 901–907

Price, W.N., Chen, Y., Handelman, S.K., Neely, H., Manor, P., Karlin, R., Nair, R., Liu, J., Baran, M., Everett, J.K., Tong, S.N., Forouhar, F., Swaminathan, S.S., Acton, T.B., Xiao, R., Luft, J.R., Lauricella, A., DeTitta, G.T., Rost, B., Montelione, G.T. & Hunt, J.F. (2009). Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nature biotechnology* 27 (1) pp. 51–57

Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic acids research* 37 (Database issue) pp. D32–36

Pusey, M.L., Liu, Z.-J., Tempel, W., Praissman, J., Lin, D., Wang, B.-C., Gavira, J.A. & Ng, J.D. (2005). Life in the fast lane for protein crystallization and X-ray crystallography. *Progress in biophysics and molecular biology* 88 (3) pp. 359–386

Quinlan, J.R. (1993). *C4.5: programs for machine learning* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Radzicka, A. & Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* 27 (5) pp. 1664 – 1670

Redecke, L., Nass, K., DePonte, D.P., White, T.A., Rehders, D., Barty, A., Stellato, F., Liang, M., Barends, T.R.M., Boutet, S., Williams, G.J., Messerschmidt, M., Seibert, M.M., Aquila, A., Arnlund, D., Bajt, S., Barth, T., Bogan, M.J., Caleman, C., Chao, T.-C., Doak, R.B., Fleckenstein, H., Frank, M., Fromme, R., Galli, L., Grotjohann, I., Kirian, R.A., Koopmann, R., Kupitz, C., Lomb, L., Martin, A. V, Hunter, M.S., Johansson, L.C., Kassemeyer, S., Katona, G., Mogk, S., Neutze, R., Shoeman, R.L., Steinbrener, J., Timneanu, N., Wang, D., Weierstall, U., Zatsepin, N.A., Spence, J.C.H., Fromme, P., Schlichting, I., Duszenko, M., Betzel, C. & Chapman, H.N. (2013). Natively inhibited Trypanosoma brucei cathepsin B structure determined by using an X-ray laser. *Science New York NY* 339 (6116) pp. 227–230

Robson, B. & Osguthorpe, D.J. (1979). Refined models for computer simulation of protein folding. *Journal of Molecular Biology* 132 (1) pp. 19–51

Rodrigues, A.P.C. & Hubbard, R.E. (2003). Making decisions for structural genomics. *Briefings in bioinformatics* 4 (2) pp. 150–167

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. & Dunker, A.K. (2001). Sequence complexity of disordered protein. *Proteins* 42 (1) pp. 38–48

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering* 12 (2) pp. 85–94

Roth, B.L. (2005). Receptor systems: will mining the receptorome yield novel targets for pharmacotherapy? *Pharmacology & therapeutics* 108 (1) pp. 59–64

Sanger, F. (1949). The terminal peptides of insulin. *The Biochemical journal* 45 (5) pp. 563–574

Sanger, F., Nicklen, S. & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74 (12) pp. 5463–5467

Schaefer, R.L., Roi, L.D. & Wolfe, R.A. (1984). A ridge logistic estimator. *Communications in Statistics - Theory and Methods* 13 (1) pp. 99–113

Schwartz, T.W. & Hubbell, W.L. (2008). Structural biology: A moving story of receptors. *Nature* 455 (7212) pp. 473–474

Service, R.F. (2008). Structural biology. Protein structure initiative: phase 3 or phase out. *Science* 319 (5870) pp. 1610–1613

Service, R.F. (2005). Structural biology. Structural genomics, round 2. *Science* 307 (5715) pp. 1554–1558

Simon, Z. (1976). *Quantum Biochemistry and Specific Interactions* Tunbridge Wells, Kent, England: Abacus Press.

Slabinski, L., Jaroszewski, L., Rodrigues, A.P.C., Rychlewski, L., Wilson, I.A., Lesley, S.A. & Godzik, A. (2007a). The challenge of protein structure determination - lessons from structural genomics. *Protein science : a publication of the Protein Society* 16 (11) pp. 2472–2482

Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I.A., Lesley, S.A. & Godzik, A. (2007b). XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* 23 (24) pp. 3403–3405

Smialowski, P., Doose, G., Torkler, P., Kaufmann, S. & Frishman, D. (2012). PROSO II - a new method for protein solubility prediction. *The FEBS journal* 279 (12) pp. 2192–2200

Smialowski, P., Martin-Galiano, A.J., Mikolajka, A., Girschick, T., Holak, T.A. & Frishman, D. (2007). Protein solubility: sequence based prediction and experimental verification. *Bioinformatics* 23 (19) pp. 2536–2542

Smialowski, P., Schmidt, T., Cox, J., Kirschner, A. & Frishman, D. (2006). Will my protein crystallize? A sequence-based predictor. *Proteins* 62 (2) pp. 343–355

Smith, J.L., Fischetti, R.F. & Yamamoto, M. (2012). Micro-crystallography comes of age. *Current opinion in structural biology* 22 (5) pp. 602–612

Stevens, R.C. (2000). High-throughput protein crystallization. *Current Opinion in Structural Biology* 10 (5) pp. 558–563

Structural Genomics Consortium, China Structural Genomics Consortium, Northeast Structural Genomics Consortium, Gräslund, S., Nordlund, P., et al. (2008). Protein production and purification. *Nature Methods* 5 (2) pp. 135–146

Sumner, J.B. (1926). The isolation and crystallization of the enzyme urease. Preliminary paper. *J. Biol. Chem.* 69 (2) pp. 435–441

Sweet, R.M. & Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of molecular biology* 171 (4) pp. 479–488

Terwilliger, T.C., Waldo, G., Peat, T.S., Newman, J.M., Chu, K. & Berendzen, J. (1998). Class-directed structure determination: foundation for a protein structure initiative. *Protein science : a publication of the Protein Society* 7 (9) pp. 1851–1856

Tomii, K. & Kanehisa, M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein engineering* 9 (1) pp. 27–36

UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* 40 (Database issue) pp. D71–5

Vapnik, V.N. (1995). *The nature of statistical learning theory.* New York, NY, USA: Springer-Verlag, Inc.

Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). Completeness in structural genomics. *Nature structural biology* 8 (6) pp. 559–566

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. & Jones, D.T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20 (13) pp. 2138–2139

Waugh, D.S. (2005). Making the most of affinity tags. *Trends in biotechnology* 23 (6) pp. 316–320

Wertz, D.H. & Scheraga, H.A. (1978). Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules* 11 (1) pp. 9–15

Wilce, M.C.J., Aguilar, M.-I. & Hearn, M.T.W. (1995). Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Analytical Chemistry* 67 (7) pp. 1210–1219

Wilkinson, D.L. & Harrison, R.G. (1991). Predicting the Solubility of Recombinant Proteins in Escherichia coli. *Nature biotechnology* 9 (5) pp. 443–448

Williamson, A.R. (2000). Creating a structural genomics consortium. *Nature structural biology* 7 Suppl (1999) p. 953

Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS journal* 275 (1) pp. 1–21

Wolf, Y.I., Grishin, N. V & Koonin, E. V (2000). Estimating the number of protein folds and families from complete genome data. *Journal of molecular biology* 299 (4) pp. 897–905

Wootton, J.C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry* 17 (2) pp. 149–163

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J. & Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems* 14 (1) pp. 1–37

Yutani, K., Ogasahara, K., Tsujita, T. & Sugino, Y. (1987). Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. *Proceedings of the National Academy of Sciences of the United States of America* 84 (13) pp. 4441–4444

Zolnai, Z., Lee, P.T., Li, J., Chapman, M.R., Newman, C.S., Phillips, G.N., Rayment, I., Ulrich, E.L., Volkman, B.F. & Markley, J.L. (2003). Project management system for structural and functional proteomics: Sesame. *Journal of structural and functional genomics* 4 (1) pp. 11–23

Appendix A

List of all features considered in design of PPCpred

- **AA_{AA_i}** Composition of the 20 standard amino acid (AA) types, i.e., the count divided by the sequence length, where AA_i stands for one of 20 AAs **(20 features)**
- **AA_{exp, bur}_{AA_i}** Composition of the exposed/buried AAs (count of the exposed/buried AA_i divided by the number of all ex-posed/buried residues in a given chain) **(40 features)**
- **pI** The isoelectric point **(1 feature)**
- **{AAIndex}** The average value of a given amino acids index AAIndex over the whole sequence **(64 features)**
- **{AAIndex}_{min, max}_{5,10,15,20}** The minimal/maximal average value of the amino acid index AAIndex among all sliding windows of sizes 5, 10, 15, and 20. For chains shorter than a given window size, we use the window size equal the length of the sequence. These features are motivated by (Babnigg & Joachimiak, 2010) **(64*4*2 = 512 features)**
- **{AAIndex}_{exp, bur}** The summed value of the amino acid index AAIndex for exposed/buried residues, divided by the number of exposed/buried residues in the sequence These features are motivated by (Price et al., 2009) **(64*2 = 128 features)**
- **DIS_AVG_VAL** The average value of the predicted disorder probabilities **(1 feature)**
- **DIS_SEG** Number of the predicted disorder segments **(1 feature)**
- **DIS_RES_seg{1,5,10,15,20}** Number of the predicted disorder residues in the disorder segments which are at least 1, 5, 10, 15, and 20 residues long, divided by the sequence length. For segments with at least 1 residue, this feature represents content of the predicted disorder **(5 features)**

- **DIS_avg** The average length of the predicted disorder segments divided by the sequence length **(1 feature)**
- **DIS_max** The maximal length of the predicted disorder segment divided by the sequence length **(1 feature)**
- **DIS_{exp, bur}** Number of the predicted exposed/buried disordered residues divided by the number of exposed/buried residues **(2 features)**
- **DIS_{exp, bur}_AVG_VAL** The summed value of the predicted disorder probability for the predicted exposed/buried residues divided by the number of predicted exposed/buried residues **(2 features)**
- **SS_{SS_i}_RES_seg{1,5,10,15,20}** Number of residues in the predicted coil/helix/strand segments, $SS_i \in$ {C, H, E}, which are at least 1, 5, 10, 15, and 20 residues long, divided by the sequence length. For segments with at least 1 residue, these feature represents content of the predicted coils, helices, and strands **(15 features)**
- **SS_{SS_i}_avg** The average length of the predicted SS_i segments divided by the sequence length **(3 features)**
- **SS_{SS_i}_max** The maximal length of the predicted SS_i segments divided by the sequence length **(3 features)**
- **SS_{SS_i}_AVG_VAL** The average predicted probability be in the secondary structure state SS_i **(3 features)**
- **SS_{exp, burr}_{SS_i}** Number of the predicted exposed/buried residues in the secondary structure state SS_i divided by the number of exposed/buried residues **(6 features)**
- **RSA_AVG_VAL** The average value of predicted relative solvent accessibility **(1 feature)**
- **{EXP,BUR}_RES_seg{1,5,10,15,20}** Number of the predicted exposed/buried residues in the exposed/buried segments which are at least 1, 5, 10, 15, and 20 residues long divided by the sequence length. For segments with at least 1 residue, these features represent content of the exposed/buried residues. We note that there were no predicted exposed segments with over 15 residues, and thus the corresponding two features were removed **(8 features)**.

Appendix B

Features used by PPCpred

Following tables lists all features used by PPCpred to perform predictions for the four steps of crystallization pipelines. Each table corresponds to one crystallization step.

Features used to predict propensity of protein material production

Feature name	Biserial	Brief description
WILM950101_min_5	-0.375	Minimal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 5 residues
AA_exp_Glu	0.107	Content of the predicted exposed Glu
DIS_RES_seg_15	-0.198	Content of the predicted disordered residues in segments of 15 or more residues
KIDA850101_min_5	0.099	Minimal average value of the hydrophobicity index (Kidera et al., 1985) in a window of 5 residues
WERD780104_min_5	0.088	Minimal average value of the energy index (Wertz & Scheraga, 1978) in a window of 5 residues
AA_Cys	-0.185	Composition of Cys
LAWE840101_max_20	-0.101	Maximal average value of the energy index (Lawson et al., 1984) in a window of 20 residues
YUTK870103_max_5	0.195	Maximal average value of the energy index (Yutani et al., 1987) in a window of 5 residues
RSA_REAL	-0.133	Average value of the predicted relative solvent accessibility
AA_bur_Arg	0.087	Content of the predicted buried Arg
OOBM770101_min_15	0.095	Minimal average value of the energy index (Oobatake & Ooi, 1977) in a window of 15 residues

Features used to predict propensity of protein purification

Feature name	Biserial	Brief description
AA_bur_Ser	-0.198	Content of the predicted buried Ser
GOLD730101_max_20	-0.129	Maximal average value of the hydrophobicity index (Goldsack & Chalifoux, 1973) in a window of 20 residues
BULH740101_max_10	-0.199	Maximal average value of the energy index (Bull & Breese, 1974) in a window of 10 residues
ROBB790101_exp	-0.098	Average value of the energy index (Robson & Osguthorpe, 1979) over the predicted exposed residues divided by the length of the sequence
MANP780101_min_5	0.149	Minimal average value of the hydrophobicity index (Manavalan & Ponnuswamy, 1978) in a window of 5 residues
AA_burr_Cys	-0.191	Content of the predicted buried Cys
pI	-0.181	Isoelectric point
ROBB790101_min_15	0.118	Minimal value of the energy index (Robson & Osguthorpe, 1979) in a window of 15 residues
AA_exp_Asn	-0.092	Content of the predicted exposed Asn
AA_exp_Met	0.094	Content of the predicted exposed Met

Features used to predict propensity of protein crystallization

Feature name	Biserial	Brief description
GOLD730101_min_10	0.398	Minimal average value of the hydrophobicity index (Goldsack & Chalifoux, 1973) in a window of 10 residues
DIS_SEG	-0.288	Number of the predicted disorder segments
WILM950104_max_15	-0.201	Maximal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 15 residues
EXP_RES_seg_5	-0.163	Content of the predicted exposed residues in segments of 5 or more residues
AA_exp_His	-0.207	Content of the predicted buried His
EISD860102_min_10	0.151	Minimal average value of the hydrophobicity index (Eisenberg & McLachlan, 1986) in a window of 10 residues
ROBB790101_min_15	0.241	Minimal value of the energy index (Robson & Osguthorpe, 1979) in a window of 15 residues
KIDA850101_min_5	0.149	Minimal average value of the hydrophobicity index (Kidera et al., 1985) in a window of 5 residues
WERD780103_min_5	0.278	Minimal average value of the energy index (Wertz & Scheraga, 1978) in a window of 5 residues
SWER830101_min_5	0.139	Minimal average value of the hydrophobicity index (Sweet & Eisenberg, 1983) in a window of 5 residues

Features used to predict propensity of protein diffraction quality crystallization

Feature name	Biserial	Brief description
SS_E_avg	0.192	Average length of the predicted strand segments
HOPT810101_min_10	0.151	Minimal average value of the hydrophobicity index (Hopp & Woods, 1981) in a window of 5 residues
AA_Cys	-0.206	Composition of Cys
DIS_SEG	-0.224	Number of the predicted disorder segments
GOLD730101_min_10	0.196	Minimal average value of the hydrophobicity index (Goldsack & Chalifoux, 1973) in a window of 10 residues
SIMZ760101_bur	0.123	Average value of the energy index (Simon, 1976) for the predicted buried residues divided by the length of the sequence
AA_bur_His	0.144	Content of the predicted buried His
YUTK870103_min_10	0.135	Minimal average value of the energy index (Yutani et al., 1987) in a window of 10 residues
AA_bur_Ser	-0.223	Content of the predicted buried Ser
JURD980101_min_10	0.212	Minimal average value of the hydrophobicity index (Juretić et al., 1998) in a window of 10 residues
BLAS910101_min_15	0.153	Minimal average value of the hydrophobicity index (Black & Mould, 1991) in a window of 15 residues
WILM950102_min_10	0.147	Minimal average value of the hydrophobicity index (Wilce et al., 1995) in a window of 10 residues
RADA880104_min_5	0.166	Minimal average value of the energy index (Radzicka & Wolfenden, 1988) in a window of 5 residues
RSA_AVG_VAL	-0.175	Average value of the predicted relative solvent accessibility

Appendix C

List of all features considered in design of fDETECT

Amino acid based (420 features). These features include amino acid (AA) and dipeptides compositions.

AAComposition_{AA} – composition (count) of a given AA type {AA} divided by protein's sequence length. (**20 features**)

AAComposition_{AA}_{AA} – composition (count) of a given dipeptide {AA}_{AA} divided by protein's sequence length. (**400 features**)

Amino acid group based (336 features). Features based on division of AAs into groups characterized by specific physicochemical properties, see Table 5.2. The twenty AAs are divided into three groups for each of the seven different AA characteristics representing the main clusters of the AA indices of Tomii and Kanehisaas (Tomii & Kanehisa, 1996) that were presented in (Dubchak et al., 1999).

GRComposition_{Char}_{Gr} – Composition of AAs belonging to a given group in a given characteristic divided by protein's sequence length. This feature is computed for each group {Gr} of each characteristic {Char} in Table 5.2. (7 characteristics x 3 groups = **21 features**)

GRTransition_{Char}_{Gr_1-Gr_2 or Gr_2-Gr_1; Gr_1-Gr_3 or Gr_3-Gr_1; Gr_2-Gr_3 or Gr_3-Gr_2} – frequency of occurrence of transitions between groups for a given characteristic within the input protein. We sum AA pairs that transition between different groups and divide by protein's sequence length minus 1. This feature is computed for each of the three possible transitions (Gr_1 to Gr_2 or Gr_2 to Gr_1; Gr_1 to Gr_3 or Gr_3 to Gr_1; Gr_2 to Gr_3 or Gr_3 to Gr_2) for each group characteristic {Char} in Table 5.2. (7 characteristics x 3 transitions per group = **21 features**)

GRDistribution_{Char}_{Gr}_{first, 25th%, 50th%, 75th%, last} – Position of occurrence of {first, 25th%, 50th%, 75th%, last} residue belonging to a given group {Gr} for a given characteristic divided by protein's sequence length. This feature is computed for each group {Gr} of each characteristic {Char} in Table 5.2. (7 characteristics x 3 groups x 5 position choices = **105 features**)

GRSegmentCount_{Char}_{Gr}_{1-5, 6-10, 11-15, >15} – Count of the number of short (1-5 residues)/medium (6-10 residues)/long (11-15 residues)/very long (over 15 residues) segments of AAs that are exclusively in a given group {Gr} for a given characteristic {Char} listed in Table 5.2. These counts were normalized by the total number of segments (for that group) in the input protein chain. (7 characteristics x 3 groups x 4 segment sizes = **84 features**)

GRSegmentComposition_{Char}_{Gr}_{1-5, 6-10, 11-15, >15} – the number of AAs in the input protein sequence that are in short (1-5 residues)/medium (6-10 residues)/long (11-15 residues)/very long (over 15 residues) segments of AAs that are exclusively in a given group {Gr} for a given characteristic {Char} listed in Table 5.2. These counts were normalized by the protein's sequence length. (7 characteristics x 3 groups x 4 segment sizes = **84 features**)

GRLongestSegment_{Char}_{Gr} – the length of the longest segment of AAs that are exclusively in a given group {Gr} for a given characteristic {Char} listed in Table 5.2 divided by the protein's sequence length. This feature is computed for each group {Gr} of each characteristic {Char} in Table 1. (7 characteristics x 3 groups = **21 features**)

Amino acid index based (448 features). These features utilize per AA values of 64 hydrophobicity and energy based indices collected from the AAIndex database (Kawashima et al., 2008); The same indices that were used in PPCpred.

AAindex_{Index}_avg –average value of a given AA index {Index} over the whole input protein sequence. These features are computed for each index {Index} in Table 2. (64 indices = **64 features**)

AAindex_{Index}_{min,max}_{5, 10, 15} –The minimal/maximal average value of a given AA index {Index} among all sliding windows of sizes 5, 10, and 15 over the input protein chain. For chains shorter than a given window size, we use the

window size equal the length of the sequence. (64 indices x 6 values per index = **384 features**)

Protein's properties based (4 features). Features based on four physicochemical properties of proteins.

pI – The isoelectric point of the input protein. (**1 feature**)

AliphaticIndex – The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins. The aliphatic index of a protein is calculated according to the (Ikai, 1980). (**1 feature**)

InstabilityIndex – The instability index provides an estimate of the stability of a given protein in a test tube, with higher values denoting instable proteins with shorter *in vivo* half-life, (Guruprasad et al., 1990). (**1 feature**)

NetCharge – protein's net charge. (**1 feature**)

Disorder and complexity predictions based (75 features). These features are computed from the predictions of disordered residues performed with IUpred (Dosztányi et al., 2005), which includes predictions of both Short (IUpred_S) and Long (IUpred_L) disorder segments, and based on assignment of sequence complexity utilizing the SEG algorithm (Wootton & Federhen, 1993):

PRprobability_{IUpredL, IUPredS, Complexity}_avg – average value of probabilities/complexity values of a given predictor/algorithm {IUpredL, IUPredS, Complexity} over the whole protein sequence. (3 predictors = **3 features**)

PRprobability_{IUpredL, IUPredS, Complexity}_{min,max}_{5,10,15} – The minimal/maximal average value of probabilities/complexity values of a given predictor/algorithm {IUpredL, IUPredS, Complexity} among all sliding windows of sizes 5, 10, and 15. For chains shorter than a given window size we use the window size equal the length of the sequence. (3 predictors x 6 values per index = **18 features**)

PRSegmentCount_{IUpredL, IUPredS, Complexity}_{0, 1}_{1-5, 6-10, 11-15, >15} – count of the number of short (1-5 residues)/medium (6-10 residues)/long (11-15

residues)/very long (over 15 residues) segments in the input protein for each binary prediction/complexity value {0, 1} of each predictor/algorithm {IUpredL, IUPredS, Complexity}. These counts were normalized by the total number of segments (for that predictor) in the protein. (3 predictors/algorithm x 2 predictions/assignments per predictor/algorithm x 4 segment sizes = **24 features**)

PRSegmentComposition_{IUpredL, IUPredS, Complexity}_{0, 1}_{1-5, 6-10, 11-15, >15} – count of the number of AAs in the input protein sequence that are in short (1-5 residues)/medium (6-10 residues)/long (11-15 residues)/very long (over 15 residues) segments for each binary prediction/complexity value {0, 1} of each predictor/algorithm {IUpredL, IUPredS, Complexity}. These counts were normalized by the length of the protein. (3 predictors/ algorithm x 2 predictions/assignment per predictor/algorithm x 4 segment sizes = **24 features**)

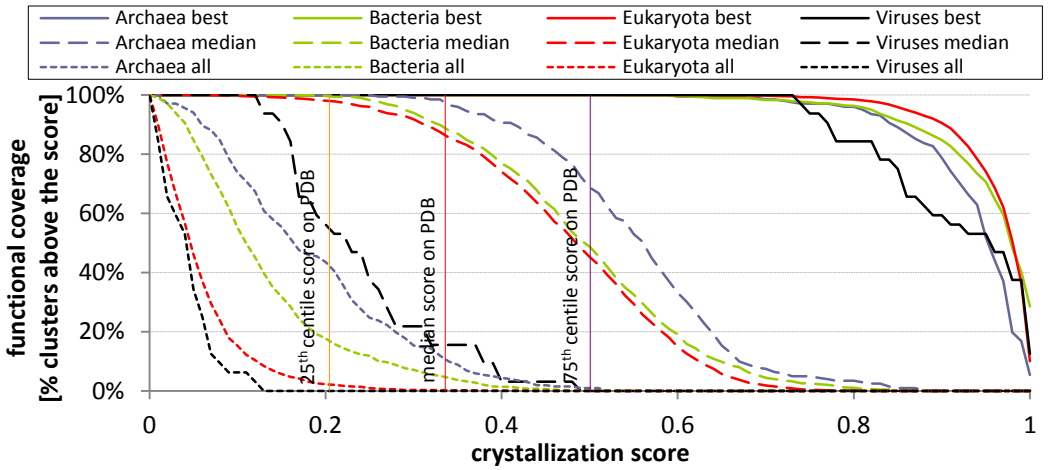
PRLongestSegment_{IUpredL, IUPredS, Complexity}_{0, 1} – the length of the longest segment for each binary prediction/complexity value {0, 1} of each predictor/algorithm {IUpredL, IUPredS, Complexity} divided by the protein sequence length. (3 predictors x 2 predictions per predictor = **6 features**)

Appendix D

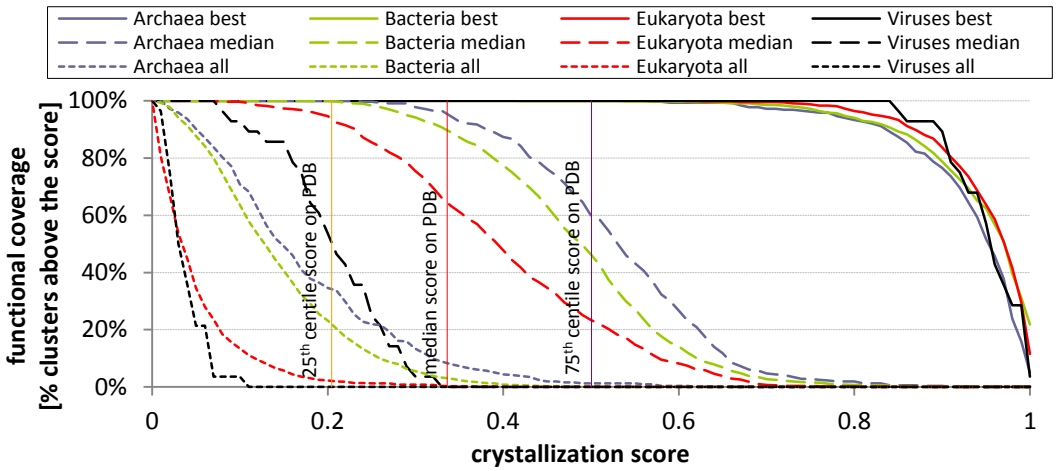
Attainable structural coverage of GO annotations

The database was clustered at 30% sequence identity and each cluster is considered solved when the crystallization score is above a given cutoff. Panels A, B, and C correspond to GO annotations concerning biological processes, molecular functions, and cellular components, respectively. The lines show the functional coverage (fraction of solved GO annotations among all available GO annotations in a given superkingdom) for a given value on the x-axis that defines the cutoff on the crystallization propensity score. To assure statistically sound estimates we limited analysis to the annotations with at least 20 clusters. The thick solid lines represent coverage where a given annotation is assumed covered when one of its annotated clusters has score above the cut-off (at least one cluster for a given annotation can be solved). The dashed lines represent coverage where a given annotation is assumed covered when at least 50% of clusters in this annotation are covered; the dotted line is when all clusters in a given annotation are covered (the annotation is fully covered).

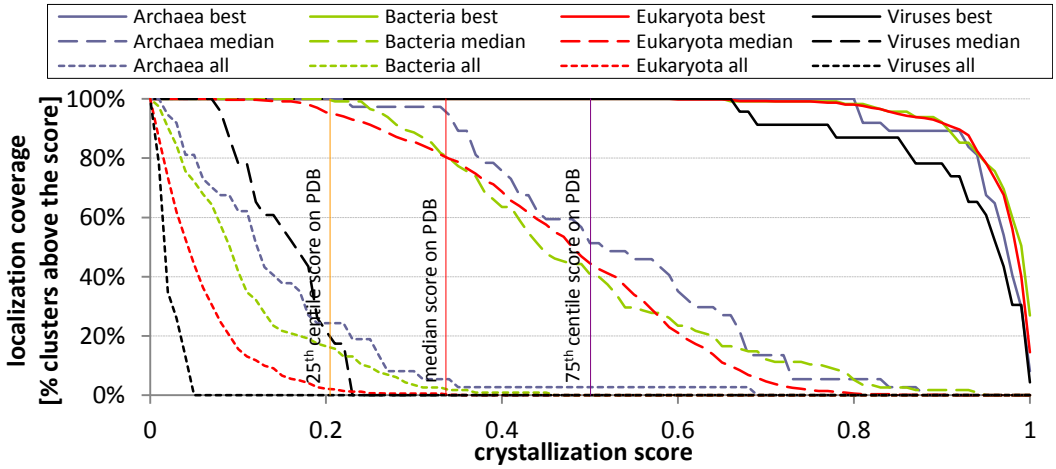
A) Biological process



B) Molecular function



C) Cellular component



Appendix E

List of high scoring GPCRs

List of G-Protein Coupled Receptors (GPCRs) with the highest predicted crystallization propensities.

UniProt ID	Protein name	Organism	Score
B0WAX1	Olfactory receptor	Culexquinquefasciatus (Southern house mosquito)	.435
D6WG06	Gustatory receptor 30	Triboliumcastaneum (Red flour beetle)	.415
Q7PK67	AGAP009706-PA	Anopheles gambiae (African malaria mosquito)	.385
B4QNG4	GD12418	Drosophila simulans (Fruit fly)	.366
Q9VVF3	Putative odorant receptor 74a	Drosophila melanogaster (Fruit fly)	.365
D6WC28	Gustatory receptor 203	Triboliumcastaneum (Red flour beetle)	.344
D6WXW3	Gustatory receptor 191	Triboliumcastaneum (Red flour beetle)	.341
B0XLS6	Odorant receptor 83c	Culexquinquefasciatus (Southern house mosquito)	.335
H0XVV2	N/A	Otolemurgarnettii (Small-eared galago)	.333
H0Y166	N/A	Otolemurgarnettii (Small-eared galago)	.329
B4H6K7	GL15497	Drosophila persimilis (Fruit fly)	.325
Q2LZG6	Or74a	Drosophila pseudoobscura pseudoobscura (Fruit fly)	.324
G3IPA0	Vomeronasal type-2 receptor 26	Cricetulusgriseus (Chinese hamster)	.322
G3U577	N/A	Loxodontaafricana (African elephant)	.319
G1U260	N/A	Oryctolaguscuniculus (Rabbit)	.315
Q7PSF0	AGAP009394-PA	Anopheles gambiae (African malaria mosquito)	.312
B3NLE6	GG21080	Drosophila erecta (Fruit fly)	.307
F7CRM9	N/A	Ornithorhynchusanatinus (Duckbill platypus)	.306
Q29H44	Or9a	Drosophila pseudoobscura pseudoobscura (Fruit fly)	.305
B0XGA0	Odorant receptor 83c	Culexquinquefasciatus (Southern house mosquito)	.305
B0VXA6	Olfactory receptor 599	Callithrixjacchus (White-tufted-ear marmoset)	.305
D6W8K5	Gustatory receptor 165	Triboliumcastaneum (Red flour beetle)	.303
B4N5C3	GK20347	Drosophila willistoni (Fruit fly)	.302
Q17NP3	AAEL000614-PA	Aedesaegypti (Yellowfever mosquito)	.302
B4GD96	GL11721	Drosophila persimilis (Fruit fly)	.301