UPPSALA
UNIVERSITET

UPTEC X 15 039
Examensarbete 30 hp
Januari 2016

Searching for novel protein-protein specificities using a combined approach of sequence co-evolution and local structural equilibration

Olle Nordesjö

UPPSALA
UNIVERSITET

Degree Project in Bioinformatics

Masters Programme in Molecular Biotechnology Engineering,
Uppsala University School of Engineering

UPTEC X 15 039 | **Date of issue 2016-01**

Author

Olle Nordesjö

Title (English)

Searching for novel protein-protein specificities using a combined approach of sequence co-evolution and local structural equilibration

Abstract

Greater understanding of how we can use protein simulations and statistical characteristics of biomolecular interfaces as proxies for biological function will make manifest major advances in protein engineering. Here we show how to use calculated change in binding affinity and coevolutionary scores to predict the functional effect of mutations in the interface between a Histidine Kinase and a Response Regulator. These proteins participate in the Two-Component Regulatory system, a system for intracellular signalling found in bacteria. We find that both scores work as proxies for functional mutants and demonstrate a ~30 fold improvement in initial positive predictive value compared with choosing randomly from a sequence space of 160 000 variants in the top 20 mutants. We also demonstrate qualitative differences in the predictions of the two scores, primarily a tendency for the coevolutionary score to miss out on one class of functional mutants with enriched frequency of the amino acid threonine in one position.

Keywords

Biotechnology, Bioinformatics, Molecular Structure, Protein Conformation, Computing Methodologies, Protein-protein interaction, Binding affinity, Mutation analysis, Amino acid mutation, *in-silico* binding affinity prediction, Two-component regulatory system, Histidine Kinase, Response Regulator, Prediction, PhoQ, PhoP, Phosphatase

Supervisors

Samuel C. Flores, Assistant Professor
Department of Cell- and Molecular Biology, Uppsala University

Scientific reviewer

Martin Ryberg, Assistant Professor
Department of Organismal Biology, Uppsala University

Project name	Sponsors
Language **English**	Security Delayed publication; 2017-01
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **34**

Biology Education Centre Biomedical Center Husargatan 3, Uppsala
Box 592, S-751 24 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 471 4687

Populärvetenskaplig sammanfattning

En stor utmaning i biologisk forskning idag är att bestämma vilken effekt olika mutationer har på proteinbinding. Vissa mutationer på viktiga proteiner som måste fungera för vårt välmående kan resultera i allvarliga konsekvenser, i värsta fall medfödda eller uppkomna genetiska defekter som begränsar möjligheter till ett normalt liv. Att bättre kunna förutse vilken effekt mutationer har på proteinbindning kan till exempel ge oss möjlighet att förutse vilka konsekvenser en mutation kan ha på olika hälsotillstånd och bidra med möjligheten att snabbare hitta lämpliga medicinska behandlingar.

I den här rapporten studeras ett modellsystem för proteinbindning, nämligen två typer av protein som finns i bakterier och deltar i intracellulär signallering. Vi använder oss av två olika metoder för att förutse vilka mutationer som har en destruktiv effekt på signalleringsfunktionen.

Den första metoden består i att genomföra virtuella mutationer med hjälp av datormjukvara på ett protein och sedan hitta en sannolik biologisk konformation genom att flexiblilisera området kring mutationerna och minimera den fria energin. Flexibiliseringen undviker stora energiförändringar som uppkommer när mutationerna genomförs.

Den andra metoden baseras på massiva mängder av proteinsekvensinformation från olika bakterier. Genom att räkna antal aminosyror som förekommer i olika positioner i alla sekvenser kan vi bestämma statistiska modeller för vad som händer när en aminosyra muteras till en annan, och till och med när flera aminosyror blivit muterade samtidigt. Anledningen till att det är viktigt att mutera flera aminosyror samtidigt är att två mutationer tillsammans sällan ger upphov till den förväntade summan av de enskilda mutationerna.

Slutsatserna från studien är att:

- Bindningsenergi mellan proteiner är en bestämmande faktor för bevarad funktionalitet efter mutation hos det studerade proteinsystemet

- Metoderna kan klassificera vissa av mutationerna som har neutral (icke-destruktiv) effekt

- Den strukturbaserade metoden har bättre möjlighet att urskilja mutationer som är mer olika, medan den sekvensbaserade metoden i hög grad är begränsad till att förutse effekten av vissa typer av mutationer.

- Den sekvensbaserade metoden har bättre förmåga att prediktera en neutral (icke-destruktiv) effekt av mutationerna på förmågan att överföra fosfatgrupper

Examensarbete 30hp
Civilingenjörsprogrammet i Molekylär Bioteknik
Uppsala Universitetet, Januari 2016

Acknowledgements

I would like to thank everyone at the Department of Cell and Molecular Biology (ICM) at Uppsala University for a nice atmosphere and creating a nice environment.

Special thanks to my supervisor Assistant Professor Samuel Flores and my co-worker during the summer, Masoumeh Nosrati, for discussions that can range from the physical to the metaphysical.

Thanks also to my reviewer Martin Ryberg, PhD at the Department of Organismal Biology, at Uppsala University.

I would also like to express my gratitude for collaboration with Assistant Professor Faruck Morcos at the Department of Biological Sciences, and for the invitation to collaborate on this project at UT Dallas during August 2015.

Contents

List of abbreviations

AUC	**A**rea **U**nder the ROC **C**urve
CA	**C**atalytic **A**ctivity Domain of a HK
DCA	**D**irect **C**oupling **A**nalysis
DI	**D**irect **I**nteraction
DIS	**D**irect **I**nteraction score
DHp	**D**imerization and **H**istidine **P**hosphotransfer domain of HK
FPR	**F**alse **P**ositive **R**ate
HK	**H**istidine **K**inase, component 1 of a TCS
HPCC	**H**igh **P**erformance **C**loud **C**omputing
MMB	**M**acro**M**olecule**B**uilder, software package implementing ZEMu
MSA	**M**ultiple **S**equence **A**lignment
PFAM	**P**rotein **Fam**ily database
PhoQ	The histidine kinase of the PhoQ/PhoP TCS system
PhoP	The response regulator of the PhoQ/PhoP TCS system
PPI	**P**rotein **P**rotein **I**nteraction
REC	**R**eceiver domain of a RR
RMSE	**R**oot **M**ean **S**quare **E**rror
ROC	**R**eceiver **O**perating **C**haracteristic
RR	**R**esponse **R**egulator, component 2 of a TCS
TCS	**T**wo-**C**omponent Regulatory **S**ystem
TPR	**T**rue **P**ositive **R**ate
ZEMu	**Z**one **E**quilibration of **M**utants

Chapter 1

Introduction

The use of computational methods has become a great aid in the understanding of processes and mechanisms in structural biology and have become an integral part of the research environment. Statistical models requiring large amounts of data for sufficient predictive power and simulations too computationally expensive to handle 10 years ago are routinely being used to improve the understanding of observations in widely different fields ranging from evolutionary processes to mechanisms of action in enzymatic reactions. The advances have contributed to understanding of systems now very obvious to be critical to various disease states, typical examples being the understanding of how G-protein coupled receptors, tyrosine kinases and neuronal ion channels are involved in various forms of cancer and neurodegenerative diseases. In addition, the advances have allowed for a better understanding of how we can redesign proteins in order to achieve specific outcomes, improving our understanding of principles in protein engineering.

Protein engineering holds potential in a range of different applications in industry and health, from improving efficiency and safety in the production of consumer chemicals to design of therapies and fundamental understanding of biological mechanisms. An improvement in the systematic generation of variants can aid in achieving the specific goals in protein engineering quicker and more cost-effectively.

Various types of computational methods are being specifically employed to elucidate the effect of mutation on interactions between biomolecules. Here, we focus on two different methods which both promises to do this, one sequence based and one structure based. Sequence based methods focus on the analysis of the biological sequences of biopolymers; DNA, RNA and protein, while structure based methods, not surprisingly, take protein structure into account. This is often performed using simulations in an attempt to capture dynamics which are important to function. These structural simulations can demand accurate time-integration algorithms and are often computationally very expensive. Together, these sequence based and structure based methods can be used to approach questions of functional importance with a massive temporal range, from rapid movements of interfacial amino acids in the case of structural simulations to evolutionary time for sequence methods.

This study focuses on characterizing a protein (PhoQ) with critical importance to intracellular prokaryotic signalling using a combined sequence and structure computational approach. This protein belongs to the Two-component regulatory system in which the two components are a sensor and an effector protein, between which a phosphotransfer reaction takes place. These proteins are of specific interest in protein engineering due to the very general prospects of redesigning intracellular pathways. Two proxies of the retained functionality of this transfer reaction upon mutation are characterized in order to predict the effect of mutation on the interface. Predicting the effect of mutation accurately will serve as a first step in very general surface redesign in order to remodel intracellular signalling pathways in prokaryotes in various ways, for example in redesigning specificity, modulating transient binding strength or specifically disrupting binding to one component while retaining binding to another in the case of branched pathways.

1.1 Specific Aims

This study aims to predict which of 160 000 mutated variants (representing all 20^4 possible substitutions at 4 residue positions) of a certain histidine kinase, PhoQ, remain functional in their ability to perform phosphotransfer to its Two-component signalling partner PhoP. This is performed using a sequence-based method, Direct Coupling Analysis (DCA) [1], and a structure-based method, Zone Equilibration of Mutants (ZEMu) [2], which ultimately evaluates binding affinity. Another related aim is to evaluate whether the methods have predictive biases.

1.2 Limitations of the study

The project is restricted to the evaluation of the ability to predict functionality of variants of PhoQ in the PhoQ/PhoP phosphotransfer reaction as this system has data available sufficient for the method validation. However, the results are likely to be transferrable to other systems, as both ZEMu and DCA have been validated for other purposes (eg. $\Delta\Delta G$ and structure prediction) on a variety of protein complexes in prior published work [1–5].

1.3 Thesis structure

Chapter 2 Explains the biological two-component regulatory system which was used and its biological significance. Explains the computational methods which were used and some background about their development and use in other settings. Explains the generation of the dataset on which the methods were applied.

Chapter 3 Describes the design of the computational methods as classifiers. Describes the results of using the classifiers on the dataset.

Chapter 4 Outlines the impact of the work and suggests further improvements to the method, areas of application, and current shortcomings of the method.

Chapter 2

Background

2.1 Two-component Signalling system

While there are many examples of biological signalling systems, some are more modular (in terms of component exchangeability) and reused than others. Signalling systems with exchangeable components provide ideal examples of systems which can be engineered and made fit to a wide range of purposes. The Two-component regulatory system (TCS) is one of these widely used modular signalling systems, commonly consisting of two components; a histidine kinase (HK) which senses input of various in modality (ion concentration, pressure, light, oxygen level), and a response regulator (RR) which transmits the signal to some effector mechanism, commonly transcription of a gene. This section will discuss some aspects of the TCS which are especially interesting in light of this thesis, namely the signalling mechanism, previous attempts to redesign the system, and recent efforts of exhaustively mapping the effects of mutations in a crucial interface between HK and RR.

2.1.1 The signalling mechanism in TCS

The signalling mechanism in TCS is highly modular with exchangeable parts. The HK commonly consists of two parts, an extracellular sensory domain, and a dimeric intracellular domain which contains the histidine that binds the phosphoryl group (DHp), and the domain responsible for phosphorylating the histidine, named after its catalytic activity (CA). The phosphoryl group on the histidine on the DHp domain is central for transmission of the signal. The response regulator in turn consists of a recognition domain (REC), specifically binding to the DHp domain and an effector domain. The signalling mechanisms itself consists of the sensor recognizing some input signal using the sensor domain. This activates the catalytic activity of CA, which performs the phosphorylation of DHp, resulting in a phosphorylated DHp domain. Subsequently, molecular recognition between the DHp in HK and the REC in RR allows for phosphotransfer to the REC domain. Downstream, the effector can then proceed with performing its actions. From the understanding of this mechanism stems a number of opportunities to redesign the system, notable examples being sensory domain exhange and exchange of residues responsible for HK-RR recognition in order to rewire specificity.

2.1.2 Previous attempts and limitations in redesigning HK/RR

Some progress to redesign HK/RR has been made by focusing on domain-domain exchange, an approach used in various forms of protein engineering, for example in the design of calcium indicator proteins [6–8]. Some attempts have been successful as evidenced by the demonstration of exchange between different sensor domains and thus exchange of the sensory modality of the signalling system[9, 10]. This exchange was performed on a domain basis between already naturally occuring sensory domains, and does thus not directly suggest the ease of successfully redesigning specificity between HK and RR. Another notable study on the exchange of a small number of amino acids at the sites of interaction between HK and RR has proven that exchange of crucial amino acids can indeed redesign specificity predictably [11]. There were still severe limitations in this study in that only known modules from close homologs were exchanged to demonstrate specificity redesign, and no attempt of exhaustive sequence mapping was performed. Exhaustive mapping is ultimately needed in order to fully characterize the molecular recognition between HK and RR.

2.2 Recent attempts to exhaustively characterize the interface

Other recent studies have used direct coupling analysis (DCA), a sequence-based statistical method to infer residue-residue interaction on the TCS systems in the attempt to predict specificity [3]. These results have only been validated on a limited scale due to the limited data availability. Finally, in early 2015, the characterization of retained phosphotransfer ability in a set of 160 000 mutants of the histidine kinase PhoQ, one member of the TCS family was performed [12]. The study focused on one TCS which involved the PhoQ (HK) and the PhoP (RR) proteins. This system allows for bacterial sensing and responding to magnesium concentration. Exchanging the native gene transcribed by the native promoter of the system, pmgrB, to a fluorescent reporter, allowed for screening for retained functionality in a model system of *E. coli.* Site-saturation mutagenesis in a number of sites of interest determined from co-evolution and subsequent flow cytometric profiling of the populations allowed for determination of the 1% of mutations which had retained phosphotransfer ability. This data finally provides the possibility of evaluating different computational methods to predict retained functionality as a first step to predict specificity.

2.3 Making use of co-evolution and structural protein data

The availability of large amounts of sequence data on TCS and the availability of separate crystallographic structures for the PhoQ/PhoQ system allows the integration of methods making use of both types of data. In the following section, an overview of the two different methods using this information is described.

2.3.1 Co-evolutionary information

One method of quantifying interaction in amino acid pairs between proteins has been to make use of the concept of co-evolution. The basic idea is that certain pairs of amino acids tend to occur in a dependent fashion, conditional upon their pairwise proximity in an interface [13]. Until recently, flaws in the methodology of deriving statistics from sequence data have limited the applicability of methods for generating useful data [13]. However, recent breakthroughs in disentangling so called direct interactions from indirect interactions have allowed greater predictive ability of actual physical interactions to be made [3]. This in combination with a greater availability of sequence information continues to increase the feasibility of leveraging the information provided by co-evolution sequence methods.

Information scores

One specific method to make leverage direct information to determine interaction specificity was presented in [1]. This method relies on 1) concatenating a large number of natively interacting protein pairs (say protein A and protein B) to a multiple sequence alignment (MSA), such that every concatenated sequence will represent one interaction between protein A and B with amino acids ($1 \rightarrow N_A$, where N_A is number of amino acids in protein A) from protein A and ($N_A + 1 \rightarrow N_A + N_B$) from protein B (See figure 2.1). Performing frequency counts in a *single* column in this MSA can give information about conservation, while performing frequency counts of *all* possible pairwise columns gives information about the co-presence of specific residue pairs.

Mutual information

A common measure of the co-presence of residue pairs has traditionally been given by mutual information:

$$MI(i,j) = P_{ij}(s_i, s_j) log(\frac{(P_{ij}(s_i, s_j))}{P_i(s_i)P_j(s_j)}) \tag{2.1}$$

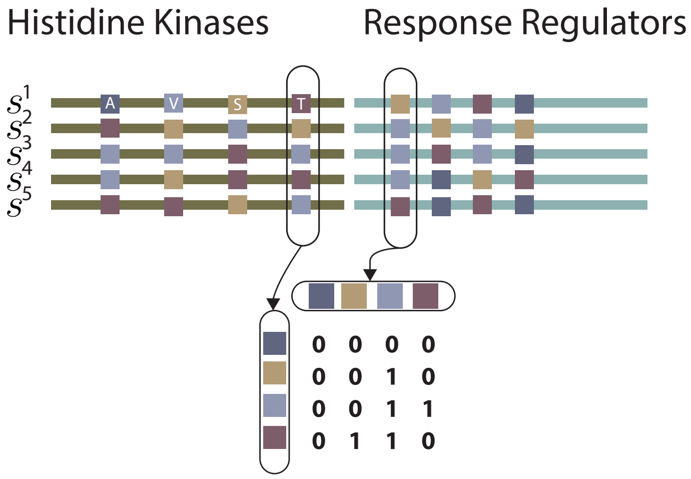
where $P_{ij}(s_i, s_j)$ represents the probability of finding position pair (i,j), (say 4 and 69) for the amino acid pair (s_i, s_j), (say alanine and proline) and $P_i(s_i)$ represents the marginal probability distribution over the amino acid s_i.

The mutual information score suffers from the problem that transitive co-presence will be included. Stated explicitly, if amino acids x and y are frequently co-occuring with a third amino acid z, then not only will the mutual information MI(x,z) and MI(y,z) be high, but MI(x,y) will also be high, even though x and y might not be in direct contact with each other. In order to eliminate this artifact, the Direct Information Score is introduced.

Direct information

The Direct Information Score (DIS) mentioned above is calculated by using methods for eliminating the unwanted indirect interaction between each inter-protein residue pair (i,j). One of these methods consists of performing a parameter optimization of the single and pair frequency counts for maximum entropy under the constraints of fulfilling marginal distribution criteria. Basically, one of the constraints is that the frequency count of a residue pair (say, $s_i = A; s_j = V$) need to be equal to the marginal distribution of the counts where $(s_i \neq A; s_j \neq V)$. Using this method of optimization, one arrives to a 4-dimensional $LxLx20x20$ probability matrix (L being the total number of amino acids in the MSA), P_{ij}^{dir}, giving direct information conditional of all LxL amino acid pairs for each position pair (i,j). Using this matrix, one can estimate the probability that a query sequence will occur: $P(A_1, ..., A_N)$ on the basis on the initial MSA. The method for deriving P_{ij}^{dir} can be found in detail in previous work [1].

In order to predict whether a query sequence will interact with a cognate sequence, the direct information score (DIS) is simply defined as the sum over all inter-protein position pairs of the sum of the DI for all possible amino acid pairs. If s represents the entire amino acid sequence, and s_i represents the amino acid at position i, the DIS is represented as:



$$\begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{matrix}$$

Figure 2.1: **A frequency count from an paired multiple sequence alignment**. A paired multiple sequence alignment is constructed, here with 5 sequences. Every sequence consists of a sequence from a HK/RR pair which have been determined to be interacting using genomic adjacency. A frequency count of the amino acid co-presence is illustrated above for the position pair (N_{HK}, N_{HK+1}). For example, in the enricled positions, we find five different residue pairs. In these pairs, the HK residues are distributed along rows of the matrix, while RR residues are distributed along columns.

$$DIS = \sum_{i \in HK, j \in RR} P_{ij}^{dir}(s_i, s_j) \frac{(P_{ij}^{dir}(s_i, s_j))}{P_i(s_i) P_j(s_j)} \tag{2.2}$$

where $P_{ij}^{dir}(s_i, s_j)$ represents the DI value at position pair (i,j) for the amino acid pair (s_i, s_j) and $P_i(s_i)$ represents the marginal distribution over the amino acid s_i.

2.3.2 Structural information

Various methods can be used for using structural information in order to say something about binding affinity. A thorough review of this subject describes the most successful ones to perform side-chain sampling, evaluation of electrostatic and solvation effects, and consideration of the monomer stability in the unbound state [14]. The method (ZEMu) used in this study was included in the evaluation.

Zone equilibration of Mutants

The requirement for evaluation of the binding affinity between the PhoQ and the PhoP was a method with the potential of performing a large number of flexibilizations rapidly for a small number of mutations. One such method is the Zone Equilbration of Mutants (ZEMu), described in earlier work [2]. This method uses an internal coordinate framework which has a number of advantages. It eases multiscale treatment by limiting the degrees of freedom of the system by means of representing positions as relative lengths and angles instead of in cartesian coordinates. It also allows for 1) specifying which bonds and angles will be flexible in the simulations and 2) specifying which atoms will have a physical influence on the flexible elements. This method has been shown to predict the $\Delta\Delta G$ due to a mutation with a RMSE of ± 1.60 kcal/mol. Zone Equlibration of Mutants is included in the software package MacroMolecule-Builder [15], which is freely available. Details on the implementation and use of this method is described in chapter 3.

Chapter 3

Evaluation

3.1 Designing the classifier

The classifier contains the two main components already mentioned in the background, the co-evolutionary approach of finding statistically significant interactions between residues, and the structural approach of minimizing interfacial energy and subsequently estimating binding affinity. This section details how the pipeline for the calculation was set up, and design choices made in the process. The two figures 3.1 and 3.2 illustrate the design of generation of the scores for all mutants.

3.1.1 DI score

The DI score was calculated using 30 000 aligned sequences from the PFAM families HIS_KA and Response_reg as detailed in previous work [3]. These pairs were assumed to be *in vivo* cognate partners, and thus contain information about the native interactions. In this paradigm, mutual information between the sites and for each amino acid 2.1 is first calculated, and subsequently a correction for indirect coupling is performed using an approach of maximum entropy. Further information about the maximum entropy can be found in the publication [3]. When this correction has been performed, the direct information matrix was used to evaluate the probability of 160 000 PhoQ mutants interacting with the still native PhoP protein, purely based on sequence data. This was done according to the following equation, described further in the background:

$$DIS = \sum_{i \in HK, j \in RR} P_{ij}^{dir}(s_i, s_j) \frac{(P_{ij}^{dir}(s_i, s_j))}{P_i(s_i) P_j(s_j)} \delta_{ij} \tag{3.1}$$

The only addition to this model is the Kronecker-delta multiplier (δ_{ij}, being

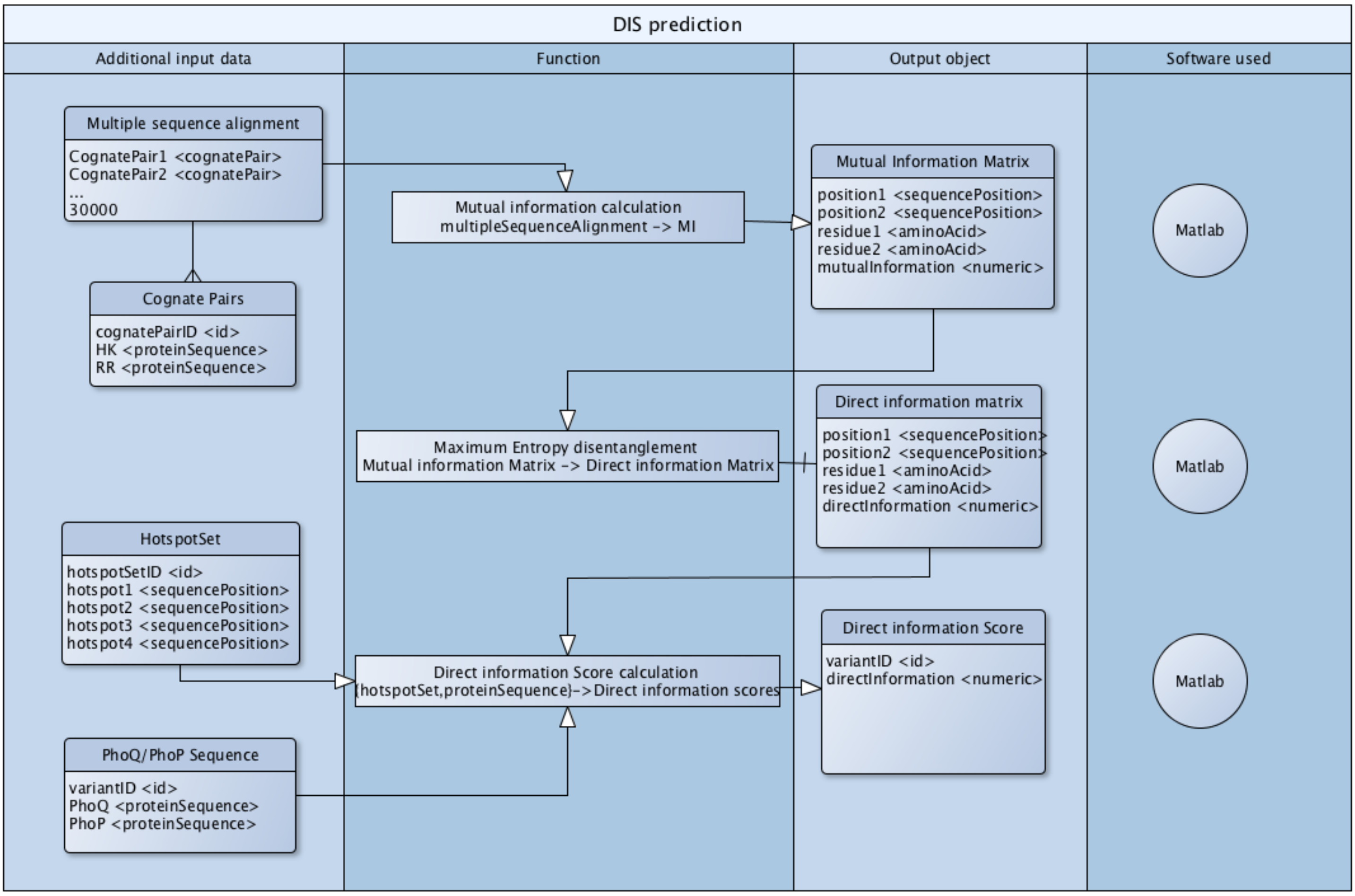

DIS prediction
Additional input data
Function
Output object
Software used
Multiple sequence alignment
CognatePair1 <cognatePair>
CognatePair2 <cognatePair>
...
30000
Cognate Pairs
cognatePairID <id>
HK <proteinSequence>
RR <proteinSequence>
HotspotSet
hotspotSetID <id>
hotspot1 <sequencePosition>
hotspot2 <sequencePosition>
hotspot3 <sequencePosition>
hotspot4 <sequencePosition>
PhoQ/PhoP Sequence
variantID <id>
PhoQ <proteinSequence>
PhoP <proteinSequence>
Mutual information calculation
multipleSequenceAlignment -> MI
Maximum Entropy disentanglement
Mutual information Matrix -> Direct information Matrix
Direct information Score calculation
(hotspotSet,proteinSequence)->Direct information scores
Mutual Information Matrix
position1 <sequencePosition>
position2 <sequencePosition>
residue1 <aminoAcid>
residue2 <aminoAcid>
mutualInformation <numeric>
Direct information matrix
position1 <sequencePosition>
position2 <sequencePosition>
residue1 <aminoAcid>
residue2 <aminoAcid>
directInformation <numeric>
Direct information Score
variantID <id>
directInformation <numeric>
Matlab
Matlab
Matlab

Figure 3.1: **Pipeline for calculating the direct information score.** The first predictor which is a co-evolutionary predictor, is determined using the scheme above. A collection of cognate pairs in the same protein family (PFAM :PF00512 and PFAM:PF00072) were used to initially determine the mutual information between position pairs in the histidine kinase vs. response regulator sequences. Following this, a method for disentangling direct interactions from indirect interactions using maximum entropy yield the direct information matrix. For each of the mutated variants, elements in this matrix corresponding to the hotspot positions are then summed to yield the direct information score, a scalar quantity with information about the probability of functional interaction between the variant and PhoP. The data and scripts for performing the initial analysis was kindly provided by Assistant Professor Faruck Morcos, UT Dallas.

1 if the amino acids (i and j) are within a distance of 12Å of another, and 0 otherwise.

3.1.2 ZEMu calculation

The pipeline for performing the binding affinity is slightly more involved. It consists of a number of steps in this case, homology modelling, declashing, *In silico* site saturation mutagenesis, equilibration, and finally energy estimation.

Homology modelling

Homology modelling was performed as there was no available co-crystal as of this writing. A co-crystal or alternatively a homology model of the complex of interest is necessary for performing the equilibration and affinity estimation. Thus, the sequences of PhoP(UNIPROT:P23836) and PhoQ(UNIPROT:P23837) were threaded to a respective substructure in a crystallised close homolog (*Thermotoga maritima*, ThkA/TrrA, PDB:3A0R) which was the one giving a structure comparable to the structures of PhoP and PhoQ in unbound form, using the MUSTER web server [16]. The one-to-one threading was performed using Phyre2, giving confidence values of > 99% for the protein to adopt the same overall fold, and a high-accuracy (2-4Å) core for both PhoQ and PhoP [17].

Declashing

As the PhoQ and PhoP were separately threaded against the close homolog and subsequently merged into a single model, a stage of residue declashing was performed in order to avoid subsequent costly structural equilibrations. The UCSF Chimera package was used to determine which residues were in clashing relation with each other using standard parameters (van der Waals overlap of 0.6 Å was considered to be a clash), followed by successive equilibrations of 15 ps using only van der Waals force fields in the MacroMoleculeBuilder package [15]. Declashing was perfomed with flexibility at the clashing sites until no more clashes were reported.

In silico site saturation mutagenesis

Subsequent to the declashing, the hotspot positions earlier determined to be especially important in the interaction were mutated to accommodate the full range of 160 000 (20^4) structural variants [11].

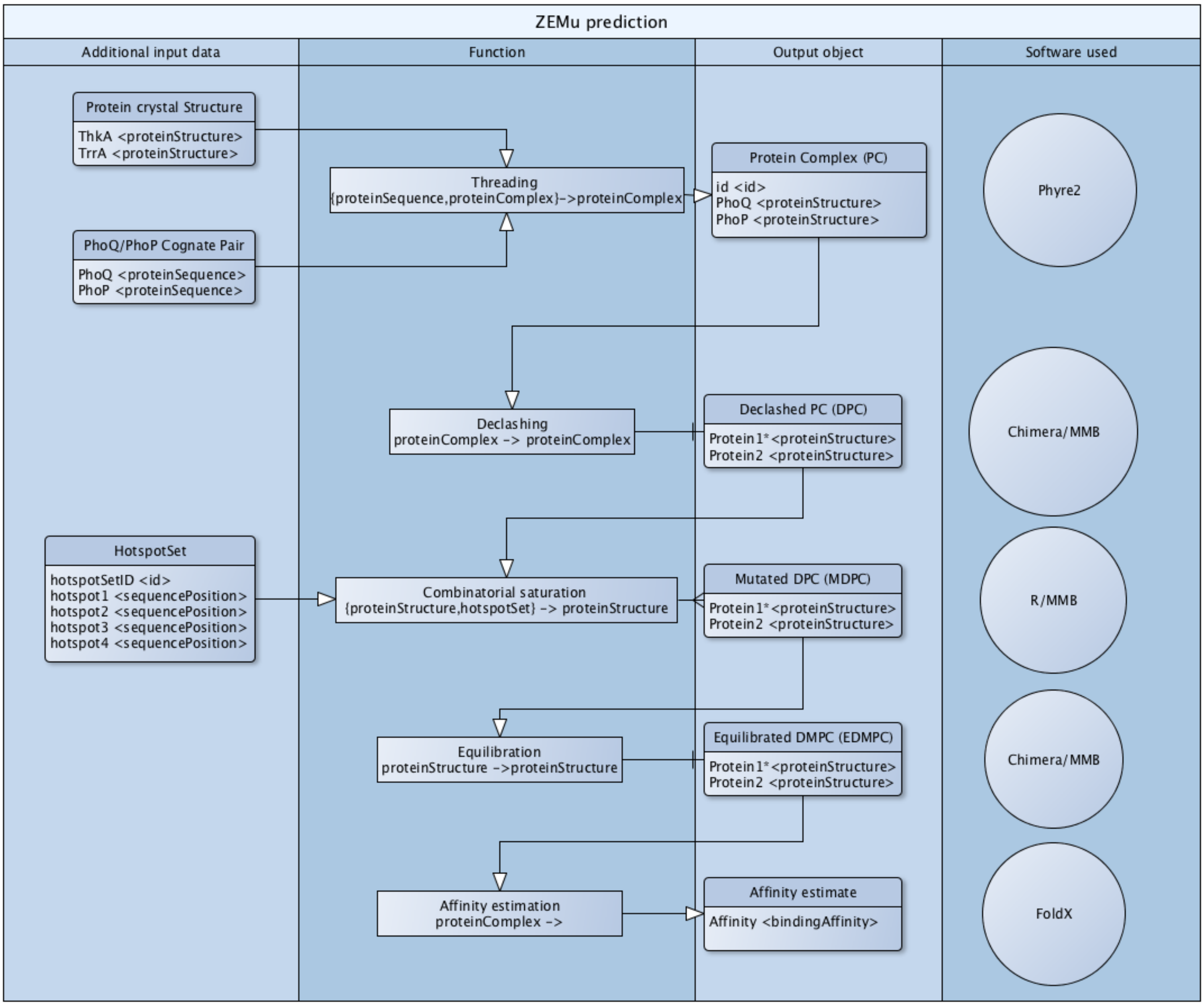

ZEMu prediction
Additional input data
Function
Output object
Software used
Protein crystal Structure
ThkA <proteinStructure>
TrrA <proteinStructure>
PhoQ/PhoP Cognate Pair
PhoQ <proteinSequence>
PhoP <proteinSequence>
HotspotSet
hotspotSetID <id>
hotspot1 <sequencePosition>
hotspot2 <sequencePosition>
hotspot3 <sequencePosition>
hotspot4 <sequencePosition>
Threading
{proteinSequence,proteinComplex}->proteinComplex
Declashing
proteinComplex -> proteinComplex
Combinatorial saturation
{proteinStructure,hotspotSet} -> proteinStructure
Equilibration
proteinStructure ->proteinStructure
Affinity estimation
proteinComplex ->
Protein Complex (PC)
id <id>
PhoQ <proteinStructure>
PhoP <proteinStructure>
Declashed PC (DPC)
Protein1*<proteinStructure>
Protein2 <proteinStructure>
Mutated DPC (MDPC)
Protein1*<proteinStructure>
Protein2 <proteinStructure>
Equilibrated DMPC (EDMPC)
Protein1*<proteinStructure>
Protein2 <proteinStructure>
Affinity estimate
Affinity <bindingAffinity>
Phyre2
Chimera/MMB
R/MMB
Chimera/MMB
FoldX

Figure 3.2: **Pipeline for calculating the ZEMu energy**. A number of methodologies are used to arrive at the final binding energy between the protein pairs. Initially, a homology model is created by threading to a close homolog. Subsequently, declashing is performed to reduce the required computation time in the following equilibration. *In silico* site saturation mutagenesis is performed at the hotspot sites which were predetermined from previous experiments. In the final steps, local equilibration and binding affinity evaluation is performed, leading to an estimate of the binding affinity, which is used as a proxy for retained functionality in the 160 000 mutants. The pipeline involved the use of a number of different software packages due to the heterogeneous requirements.

Equilibration

Equilibration of the 160 000 variants was performed using the MacroMolecule-Builder package using standard parameters: 1) Physics zone of 12 Å around mutated residues and 2) 5-consecutive-residues flexibility zone centered about each mutation position.

Binding affinity estimation

Finally, the binding affinity for each mutational complex was evaluated using the empirical force field and software package FoldX [18]. The FoldX package performs a weighted summation of a number of energy terms in order to evaluate the binding energy:

$$\begin{aligned}\Delta G = W_{vdw}\Delta G_{vdw} + W_{solvH}\Delta G_{solvH} + W_{solvP}\Delta G_{solvP} + \Delta G_{wb} + \\ \Delta G_{hbond} + \Delta G_{el} + W_{mc}T\Delta S_{mc} + W_{sc}T\Delta S_{sc}\end{aligned} \tag{3.2}$$

In the equation above, *vdw* is van der Waals contribution, *solvH* and *solvP* are solvation energies, *wb* water bridges, *el* electrostatic contributions, *mc* a backbone entropy term, and finally *sc* a side chain entropy term.

Two ΔG energies are calculated, the one for the wildtype, called ΔG_{wt}, and one for the variant, called ΔG_i. While these energies cannot be used directly, the relative binding affinity between variant i and the wildtype can be used according to the implementation of FoldX. The relative binding energy is thus calculated as:

$$\Delta\Delta G_i = \Delta G_i - \Delta G_{wt} \tag{3.3}$$

3.2 Technical details on implementation

To achieve sufficient computational power in practice, the current implementation requires ZEMu to be configured on a high-performance computer cluster (HPCC), depending on the number of mutants which have to be evaluated. In this case, the available computational power was limited to calculation of approximately 10 000 mutants within the scope of the study, using 150 000 CPU hours (core-hours) on 8-core Opteron 6220 processors running at 3 GHz. Evaluation of the direct information score can be performed on a desktop computer (2.8GHz quad-core Intel Core i5 processor, 16GB RAM) for the set of 30 000 cognate HK/RR pairs within 90 minutes. Due to the higher speed of the DIS calculation and with knowledge that the DIS is functioning as a predictor, the

ZEMu evaluation was performed for the top 10 000 DIS mutants in an attempt to further refine that predictive power.

3.3 Performance on the dataset

In this section, the binding affinities and the DIS data are presented in their relation to the state of the associated mutant (functional or nonfunctional). Subsequently, the performance of two suggested classifiers is presented and evaluated.

In order to quantitatively evaluate the classifier, the following performance metrics of the classification of functional vs. non-functional mutants are determined:

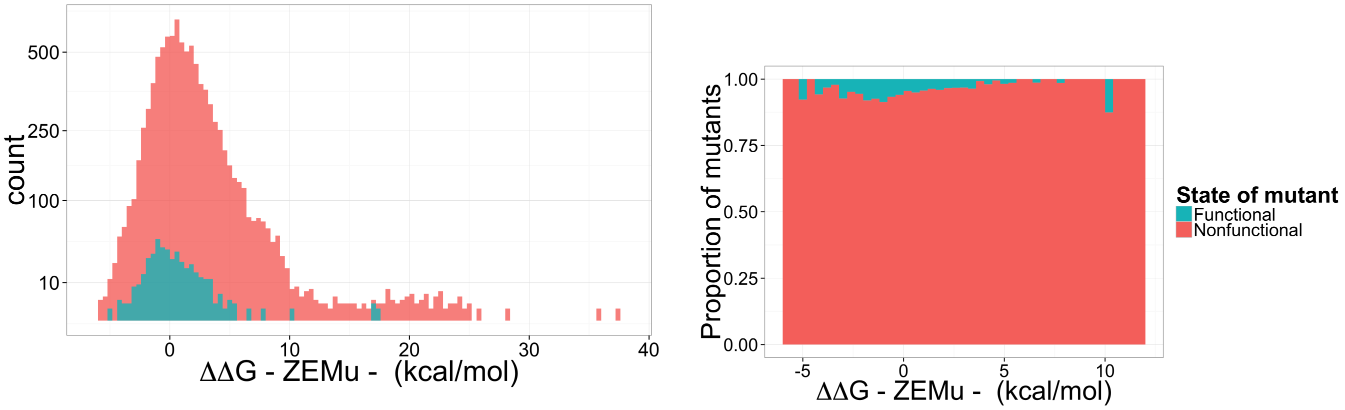
- Receiver operator characteristic of the classifier (specifically area under the curve)
- Positive likelihood ratio (TPR/FPR) of functional mutants
- Qualitative difference in classification (do the methods predict similar or different mutants?)

3.3.1 Description of the data

The change in binding affinity and the DIS have different distributions, and there is no single straight forward way of determining what is the best classification procedure to use. What is clear, however, is that a high DIS score and a ZEMu score around the mean have a higher fraction of functional mutants. A wilcoxon-rank-sum test for the difference between the means of the functional and non-functional mutants in the ZEMu data indicates that the shift between distributions is not equal to zero (p-value $< 1 \times 10^{-20}$, figure 3.3), and the difference between the means was calculated to be 1.3 ± 0.08 kcal/mol. The DIS distribution does not show a simple distribution, but the location shift for the distributions is also determined to be not equal to zero (p-value $< 2.2 \times 10^{-16}$, figure 3.4).

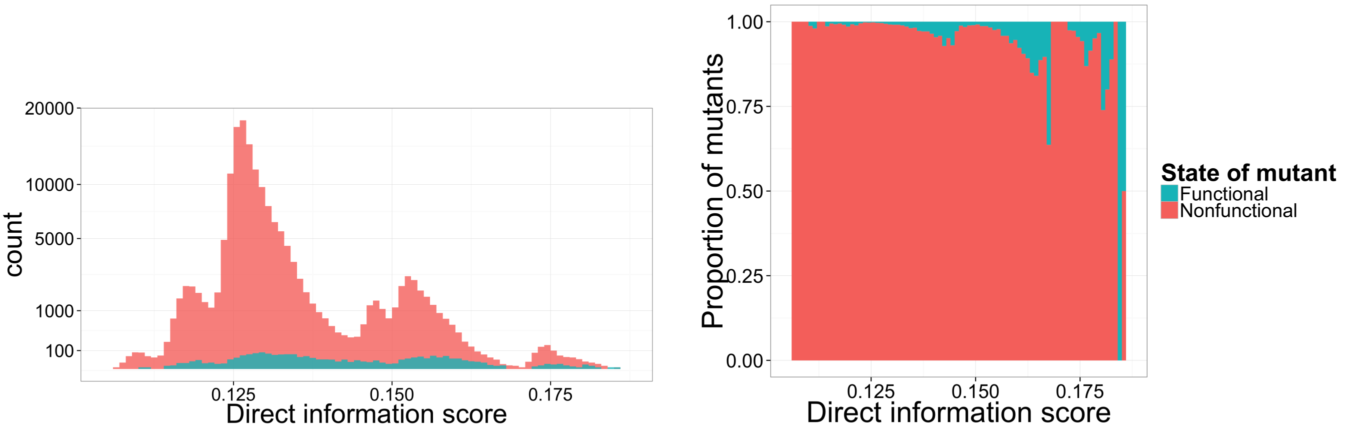
3.3.2 Receiver operating characteristic

The receiver operating characteristic for the evaluation indicates that predicting functional mutants is possible, however not completely satisfactory. The area-under-curve for the DIS only is 0.69, and for ZEMu (conditional upon the filtration to the top 10 000 DIS mutants) 0.65 (figure 3.5a). However, looking



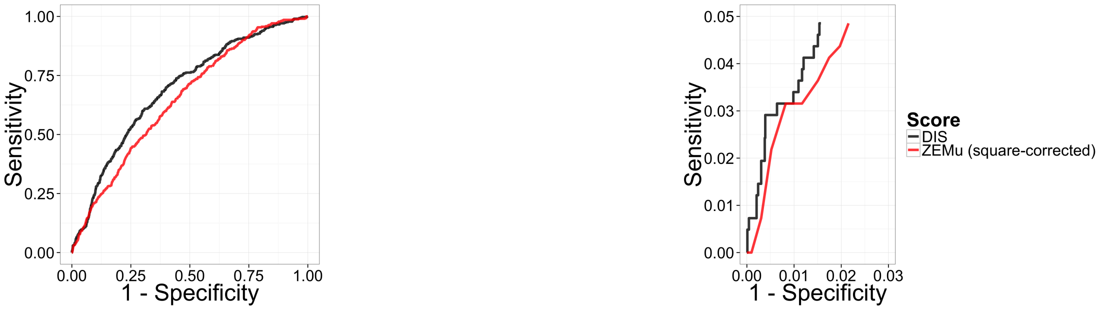
(a) (b)

Figure 3.3: **The distribution of the $\Delta\Delta G$ generated from the structural equilibration**. a) Histogram over the distribution depending on whether the mutant is functional or not. b) normalized stacked histogram for each bin, such that the proportion of the functional and nonfunctional is shown on the y-axis.



(a) (b)

Figure 3.4: **The distribution of the direct information score** . a) Histogram over the DIS distribution depending on whether the mutant is functional or not. b) Normalized stacked histogram for each bin, such that the proportion of the functional and nonfunctional is shown on the y-axis.



(a) ROC - overview

(b) ROC - detail

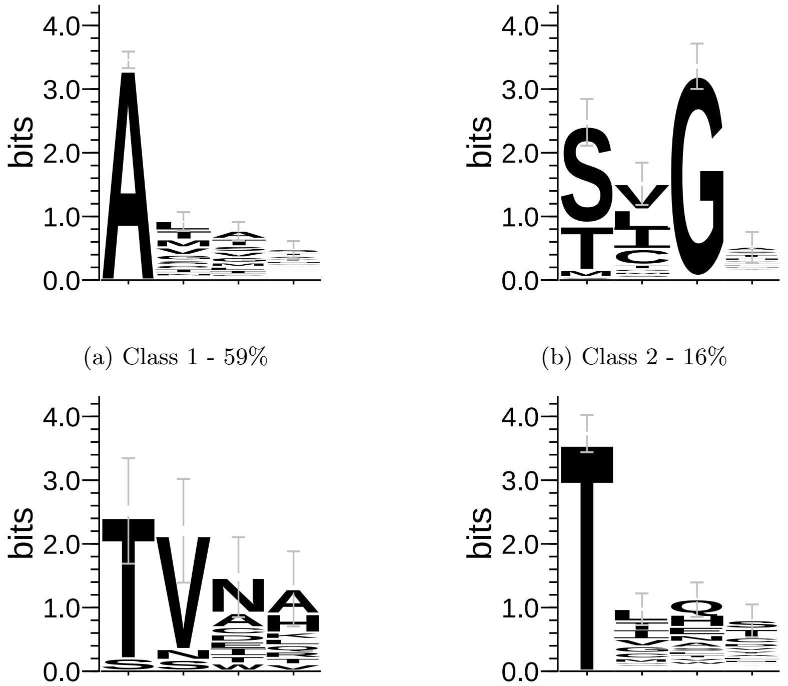
Figure 3.5: **Receiving operating characteristic for the different scores**. In general, the DI score outperforms the ZEMu score. AUC(DIS) = 0.69, AUC(ZEMu) = 0.65. However, both scores perform better than random. Note that the ZEMu scores are calculated for the mutants with top 10 000 DI score, implying that the ZEMu score can not be used in isolation.

in the lower-left corner of the curve (figure 3.5b), the maximal positive likelihood ratio value we can achieve with a number of 20 mutants is about 0.3, an improvement of 30x the prevalence of functional variants in the initial sample. This is true for both the ZEMu score and the DIS.

3.4 Evaluation of heterogeneity

In order to evaluate whether there was any heterogeneity in the classification, we developed a method for displaying to what extent the methods perform differently for certain types of mutant classes.

This was performed as follows: The functional mutants were ranked by their respective score, DIS or ZEMu score. Each mutant was assigned a class based on hierarchical clustering of all functional mutants, using the pairwise hamming string distance between the sequence of amino acids in the mutation positions as the distance score. As an example, $s_1 = AVST$, $s_2 = AVRT$ would give a string distance $dist(s_1, s_2) = 1$. Finally, the data is visualized by plotting the cumulative number of mutants within that class above a certain rank, for all ranks. Using this method, there are essentially three characteristic patterns present. First, a straight line for a certain class will indicate equal probability of that class being distributed over the range. Second, an increasing slope for lower ranks indicates less likelihood of ranking such mutants high. Third, a decreasing slope indicates a class likely to be ranked high. With this in mind, we can display how the predictors perform in figure 3.7 for the different classes, illustrated as sequence logos in figure 3.6. Adding more classes than 4 gave no further information about the difference in heterogeneity.



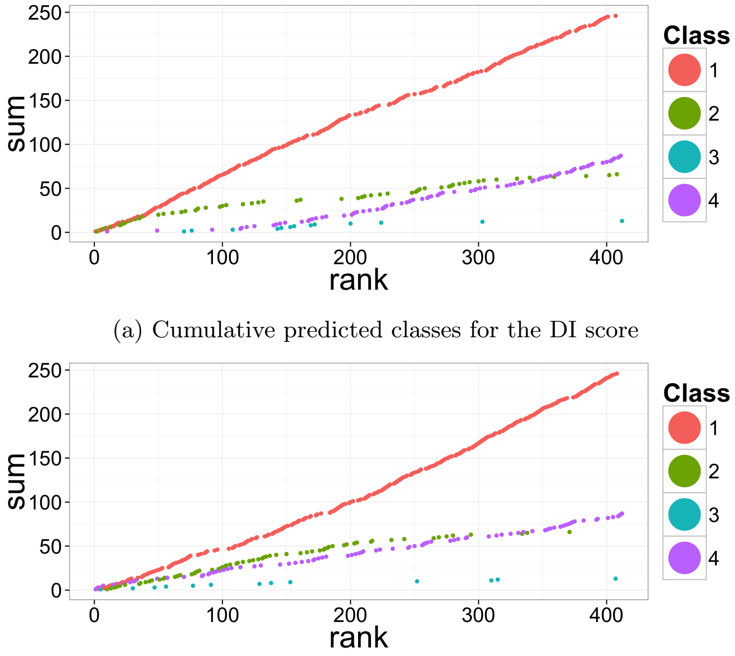
(a) Class 1 - 59%

(b) Class 2 - 16%

(c) Class 3 - 3%

(d) Class 4 - 21%

Figure 3.6: **Sequence logos of the different classes generated using hierarchical clustering of the functional variants**. Notice how class 1 and 4 are single amino acid profiles, while class 2 and 3 are multiple profiles. Notice also the varying prevalence, given by the percentage number in the captions.



(a) Cumulative predicted classes for the DI score

(b) Cumulative predicted classes for the ZEMu score

Figure 3.7: **Score heterogeneity for the different scores**. A line with constant slope indicates naivety to rank within the class, while a line with variable slope indicates that there is a preference (decreasing slope) or dispreference (increasing slope) over lower rank.

Chapter 4

Discussion and conclusions

4.1 Discussion

The previous section indicated the feasibility of predicting functional variants of the histidine kinase PhoQ based on integration of two methods, co-evolution and Zone equilibration of Mutants. The difference between the means of the $\Delta\Delta G$ values in the functional and nonfunctional populations suggest the usefulness of the change in binding energy as a proxy for functionality, which is of major importance. Succinctly stated, this indicates the fundamental result that binding affinity is a predictor of functionality in the phosphotransfer reaction. It is possible to reach a positive likelihood ratio of about 30x the initial prevalence of functional mutants using the DI score, and the ZEMu score indicates a similar performance, however with the caveat that only the mutants performing best under the DI score could be calculated due to limitations in computational time.

It is also clear that there is a heterogeneity in the types of mutants which are predicted by the different methods. The DI score, by design, favors patterns already found in the co-evolutionary profile leaving out one class prevalent in the functional variants, while ZEMu is naive to the profile and also captures functional variants of all classes similarly well. Both methods capture the quite simple pattern with simply an alanine in the first position.

These results provide valuable insights in the predictability of protein protein interactions, specifically in the case of TCS systems. However, any protein-protein complex with a similar amount of sequence data could be analysed with the co-evolutionary method, and any system where a co-crystal of a close homolog is available could make use of the ZEMu approach. There are however limitations to the speed of calculation of the mutants. This could be improved in a number of ways. One way is to enforce a dynamic convergence criterion for ZEMu, which can result in quicker convergence time of the simulations. Another attempt to improve on the results for ZEMu calculation could be to evaluate effects of different methods of protein protein docking on the ZEMu performance.

The presented results also allow for further discovery of additional suspected functional mutants, especially as the validation study [11] estimated a 7% false negative rate in discovery of the functional mutants.

4.2 Further work

Further research in this area should be focused on improving the calculation speed of the structural equilibration, validating newly found potential functional mutants, and also evaluating the applicability on other systems to determine the generality of the approach.

4.3 Conclusions

This study has successfully shown that two different methods, co-evolution with a DI score, and strucural equilibration using a ZEMu score, are applicable in predicting functional variants of the histidine kinase PhoQ. In addition, of biological importance is the fact that there is a significant difference between the population means (functional vs. nonfunctional) of the change in binding energies upon mutation for the functional and nonfunctional variants. This suggests that binding affinity predicts functionality for the phosphotransfer reaction. The methods show an AUC of 0.69 (DIS) and 0.65 (ZEMu). The positive likelihood ratio of finding a positive mutant within the top 20 predicted mutants is 0.3, a 30X increase compared to the prevalence in the complete dataset.

Bibliography

[1] Ryan R. Cheng, Faruck Morcos, Herbert Levine, and José N. Onuchic. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of Sciences of the United States of America*, 111(5):E563–E571, February 2014.

[2] Daniel F. A. R. Dourado and Samuel Coulbourn Flores. A multiscale approach to predicting affinity changes in protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2681–2690, October 2014.

[3] Faruck Morcos, Terence Hwa, José N. Onuchic, and Martin Weigt. Direct coupling analysis for protein contact prediction. *Methods in molecular biology (Clifton, N.J.)*, 1137:55–70, 2014.

[4] Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. PNAS Plus: Direct-coupling analysis of residue coevolution captures native contacts across many protein families. 108(49):E1293–E1301, 2011.

[5] Faruck Morcos, Biman Jana, Terence Hwa, and José N. Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, 110(51):20533–20538, December 2013.

[6] Aleksandra Badura, Xiaonan Richard Sun, Andrea Giovannucci, Laura A. Lynch, and Samuel S.-H. Wang. Fast calcium sensor proteins for monitoring neural activity. *Neurophotonics*, 1(2), 2014. undefined.

[7] Akira Muto, Masamichi Ohkura, Gembu Abe, Junichi Nakai, and Koichi Kawakami. Real-Time Visualization of Neuronal Activity during Perception. *Current Biology*, 23(4):307–311, February 2013. WOS:000315178400021.

[8] John G. Partridge. Utilizing GCaMP transgenic mice to monitor endogenous G(q/11)-coupled receptors. *Frontiers in Pharmacology*, 6:UNSP 42, March 2015. WOS:000352839500001.

[9] Andreas Möglich, Rebecca A. Ayers, and Keith Moffat. Design and Signaling Mechanism of Light-Regulated Histidine Kinases. *Journal of Molecular Biology*, 385(5):1433–1444, February 2009.

[10] Andreas Möglich, Rebecca A. Ayers, and Keith Moffat. Addition at the Molecular Level: Signal Integration in Designed Per–ARNT–Sim Receptor Proteins. *Journal of Molecular Biology*, 400(3):477–486, July 2010.

[11] Jeffrey M. Skerker, Barrett S. Perchuk, Albert Siryaporn, Emma A. Lubin, Orr Ashenberg, Mark Goulian, and Michael T. Laub. Rewiring the Specificity of Two-Component Signal Transduction Systems. *Cell*, 133(6):1043–1054, June 2008.

[12] Anna I. Podgornaia and Michael T. Laub. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677, February 2015.

[13] Thomas A. Hopf, Charlotta P. I. Schärfe, João P. G. L. M. Rodrigues, Anna G. Green, Oliver Kohlbacher, Chris Sander, Alexandre M. J. J. Bonvin, and Debora S. Marks. Sequence co-evolution gives 3d contacts and structures of protein complexes. *eLife*, 3:e03430, November 2014.

[14] Rocco Moretti, Sarel J. Fleishman, Rudi Agius, Mieczyslaw Torchala, Paul A. Bates, Panagiotis L. Kastritis, João P. G. L. M. Rodrigues, Mikaël Trellet, Alexandre M. J. J. Bonvin, Meng Cui, Marianne Rooman, Dimitri Gillis, Yves Dehouck, Iain Moal, Miguel Romero-Durana, Laura Perez-Cano, Chiara Pallara, Brian Jimenez, Juan Fernandez-Recio, Samuel Flores, Michael Pacella, Krishna Praneeth Kilambi, Jeffrey J. Gray, Petr Popov, Sergei Grudinin, Juan Esquivel-Rodríguez, Daisuke Kihara, Nan Zhao, Dmitry Korkin, Xiaolei Zhu, Omar N. A. Demerdash, Julie C. Mitchell, Eiji Kanamori, Yuko Tsuchiya, Haruki Nakamura, Hasup Lee, Hahnbeom Park, Chaok Seok, Jamica Sarmiento, Shide Liang, Shusuke Teraguchi, Daron M. Standley, Hiromitsu Shimoyama, Genki Terashi, Mayuko Takeda-Shitaka, Mitsuo Iwadate, Hideaki Umeyama, Dmitri Beglov, David R. Hall, Dima Kozakov, Sandor Vajda, Brian G. Pierce, Howook Hwang, Thom Vreven, Zhiping Weng, Yangyu Huang, Haotian Li, Xiufeng Yang, Xiaofeng Ji, Shiyong Liu, Yi Xiao, Martin Zacharias, Sanbo Qin, Huan-Xiang Zhou, Sheng-You Huang, Xiaoqin Zou, Sameer

Velankar, Joël Janin, Shoshana J. Wodak, and David Baker. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 81(11):1980–1987, November 2013.

[15] S.C. Flores, M.A. Sherman, C.M. Bruns, P. Eastman, and R.B. Altman. Fast Flexible Modeling of RNA Structure Using Internal Coordinates. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(5):1247–1257, September 2011.

[16] Sitao Wu and Yang Zhang. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, 72(2):547–556, August 2008.

[17] Lawrence A. Kelley, Stefans Mezulis, Christopher M. Yates, Mark N. Wass, and Michael J. E. Sternberg. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6):845–858, June 2015.

[18] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology*, 320(2):369–387, July 2002.