UNIVERSITY OF TROMSØ UIT

Faculty of Science and Technology
Department of Mathematics and Statistics

Model based Statistics for Protein Sequence Families

Cold adapted enzymes

Said H. Ahmed

A dissertation for the degree of
Philosophiae Doctor

July 2011

Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Tor Flå for all his guidance and help over the course of my doctorial studies. Without his guidance, support, and patience, this dissertation work would not have been finished. I would also like to thank my co-supervisor George Elvebakk for all the fruitful discussions we had about factor analysis of multidimensional data and for his assistance in reviewing the last paper of this research. My thanks goes to Steinar Thorvaldsen for all the discussions we had about cold-adapted enzymes.

I would like to acknowledge the University of Tromsø for the financial support of this study and my research stay in San Diego. I would also like to acknowledge the center for theoretical and computational chemistry (CTCC) for their generous support.

Finally, I would like to thank my family, for all their love and encouragement. Without their encouragement and understanding it would have been impossible for me to finish this work. And most of all for my loving, supportive and patient wife Muna whose faithful support during the final stages of this Ph.D. is so appreciated. I owe my loving thanks to my kids Rim and Nasser. They gave me other things to think and experience. Thank you.

Said H. Ahmed
University of Tromsø, Norway
April 2011.

Contents

Chapter 1

Introduction

In the first part of this introduction, we motivate statistical approaches to biological data analysis. We are especially concerned with basic features of temperature effects on cold adapted enzymes. We will study these features with respect to the concept of evolution applied to the functional constraints of a funneled protein energy landscape and minimum frustration. In the second part, we provide summary of the research reported in this thesis.

1.1 Motivation : biological data analysis

One of the applications of protein sequence and structural data is to construct methods to classify and characterize these data according to function. Protein sequences are invaluable sources of information for inferring evolutionary relationships between the species and hence to model the molecular mechanisms by which these species evolve. While it is presumed that homologous sequences have diverged from a common ancestral sequence through iterative molecular changes, it is not known what the ancestral sequence was; all we have to observe are the sequences from extant organisms. In a multiple sequence alignment (MSA), like the one shown in Fig. 1, it is often apparent that certain regions of a protein or specific amino acids are more highly conserved than others. This information may be suggestive of which residues are more crucial for a maintaining a protein's structure or function. It is therefore desirable, from sequence analysis, to identify amino acid sites that are responsible for functional and structural divergence.

Computational methods have been developed that allow for biological discovery based on MSA. Unlike nucleotide sequences, which are composed of four bases that are chemically rather similar, yet distinct, the alphabet of the 20 amino acids found in proteins allows for much greater diversity of structure and function, primarily because the differences in chemical makeup of these amino acid residues are more pronounced. Each residue can influence the overall physico-chemical properties of the protein because the amino acids are charged (acidic or basic), polar or hydrophobic as shown by the different colors in Fig. 1. Valencia and coworkers [33, 34] used principal component analysis (PCA) and self-organizing maps to extract sequence patterns characteristic of subfamilies. This approach has great potential for functional genomics because it is cost effective. Methods of distinguishing

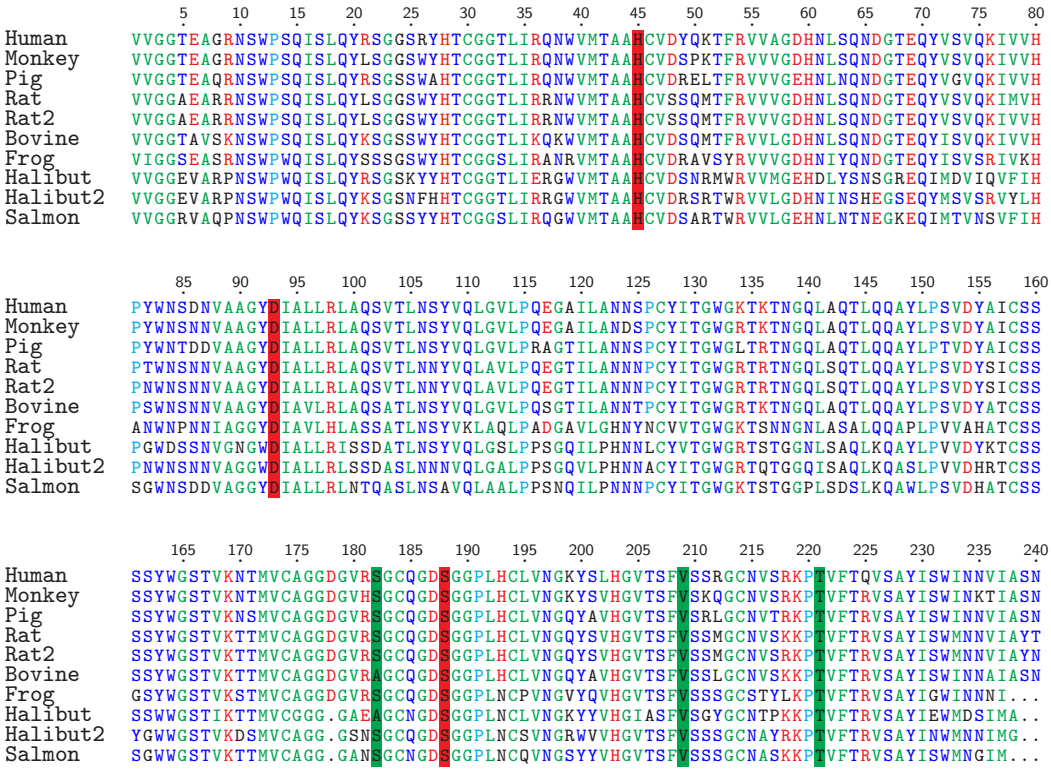

Fig. 1. Alignment of amino acid sequences of elastase type I. Gaps are shown in dots, amino acids forming the catalytic triad and the specificity pocket are shaded in red and green colors, respectively. Shown at the top is the numbering scheme of the residues as they appear in the alignment.

hydrophobic and hydrophilic regions of proteins [26] have been used to predict three-dimensional (3D) structure of proteins, their likely home within the cell, and their broad functional characteristics. It has been realized very early that protein sequences can be represented by various profiles, the most prominent one being the amino acid's hydrophobicity [26, 28], but also using other physico-chemical properties such as charge and secondary structure properties. Structural profiles, on the other hand, can also be reduced to profiles describing structural properties of the amino acids in the fold [24, 27], prominently secondary structures and solvent accessibility [25]. Recently, a 1D representation of the protein structure has been introduced [46], which is found to be related to the sequences attaining that structure via their hydrophobicity profiles [60].

A protein is capable of self-assembly and reliable functioning in a fluctuating environment. Understanding how these remarkable properties arise as the result of evolution is central to the development of a protein folding theory [3]. The energy landscape theory, formulated by Wolynes and coworkers [19, 7, 4], describes folding as occurring on a free energy surface in the shape of a rugged funnel. Therefore, protein folding can be described as a progressive organization of an ensemble of liquid partially folded states through which a protein passes on its way to folded state. They showed via spin-glass theory that there are, at least, two possible transitions : one to the folded state, which is characterized by a "folding" temperature T_f below which the folded state is thermodynamically stable, and the other to a glassy state, which is characterized by a "glass" transition temperature T_g below which the protein can be trapped in low energy misfolded states. This approach introduced the principle of minimum frustration, which asserts that nature, via evolution, has selected for sufficiently non-frustrated sequences with folded structures that are substantially stable more than other local minima on the landscape. Consequently, evolutionary selected sequences are generally thought to have globally "funneled energy landscapes" that are biased towards the folded state through any of the many pathways and intermediates. The number of pathways progressively reduce when approaching the folded state. At the bottom of the funnel the folding rate and the stability of the protein are closely related, being in some sense the kinetic and thermodynamic side of the same coin.

The minimum frustration principle and the funneling concept, therefore, suggests that the main features of folding kinetics can be predicted by knowing the stabilization energies of elements of the native structure and the entropic costs of bringing together parts of the scaffold. This design principle of the energy landscape demonstrates that the main driving force opposing the transition from the liquid-like states to the folded state is the necessary loss of conformational entropy of the protein. The mean free energy difference between the liquid protein states and the folded state is given by $\Delta F = \Delta E - T\Delta S$, where ΔE is the energy gap between the average en-

ergies in the folded state and the liquid ensemble and $\Delta S = S_0 - \Gamma_l^2/2T^2$ is the entropy at energy variance Γ_l^2 and configurational entropy of the chain, given by $S_0 = \log \Omega_0$. Here the subscript l indicates the liquid-like states, T is the absolute temperature and Ω_0 is the number of conformational states of the protein chain. A larger ΔE gives a deeper funnel. The liquid ensemble is thermodynamically stabilized by a larger configurational entropy of the disordered phase (S_0) or a large ruggedness of its energy landscape Γ_l^2. Evolution will try to avoid the misfolded disordered regime of negative entropy $T < T_g$ and unfolding regime of negative free energy difference $T > T_f$. Therefore, T_g/T_f should minimized which corresponds to minimization of the frustration parameter $\kappa^{-1} = \Gamma_l/\Delta E$ under the constraint that $\Delta F > 0$ is monitored.

We motivate a statistical approach to biological data analysis and protein folding funneling picture to parameterization of the free energy landscape difference in terms of total entropy. The main concern of this thesis will be environmental temperature effects on cold adapted enzymes. Specifically, we will be interested in the possibility of parameterizing the environment and stability effects into evolutionary models. It has been known that prediction of functional sites and site-specific amino acid distributions can be improved considerably if phylogenetic trees and evolutionary models are considered. In the evolutionary models, environment of a protein site within the folded structure is known to influence the probability of acceptance of a mutation at that site [12]. Both the mutational process and the selection on protein folding and function must be therefore taken into account. Evolution works on a rather crude energy scale, making sure the folding landscapes are robust and funnel like [3]. Thus, the structures of naturally occurring proteins are selected by evolution because they have a high sequence entropy (a high "designability").

Cold adapted enzymes

Extreme environmental temperatures are those that fall outside the limited range at which we and most other eukaryotes can survive. Organisms that are adapted to cold environments are often termed as psychrophiles, whereas their warm-active counterparts are referred to as mesophiles and thermophiles. Proteins from these organisms must be stable and functional throughout the entire range of temperatures that these organisms experience. Surprisingly, the same mechanisms of protein stabilization act on both psychrophiles at temperatures close to freezing point of water and hyperthermophiles acting at temperatures above the boiling point. Because the relatively modest thermodynamic stability of proteins (5-15 Kcal/mol) results from the small difference between large stabilizing (enthalpy) and destabilizing (conformational entropy) contributions, small changes in the number and strength of these contributions can cause a large proportional

change in this difference [20, 21]. In fact, the wide range of interactions that can be adjusted means that different proteins from different organisms can use various combinations of modulations to adjust their thermostability. This has made it difficult to delineate general rules for molecular basis of temperature adaptation [73].

Although psychrophilic organisms are distributed widely in nature, little progress has been made in elucidating a molecular basis of cold adaptation. This could be due to the low number of available sequences and structures from these enzymes. Recent accumulation of sequence and structural data from psychrophilic enzymes is beginning to shed light on the adaptation strategies of these enzymes [76]. The dominating hypotheses of cold adaptation strategy for enzymes points towards relationships between stability, flexibility, and specific activity [76, 74]. The increased flexibility of psychrophilic enzymes could be attributed to global structure, but may equally well be the result of local flexibility. The local flexibility, especially around the active site, seems to be a strategy for cold-adapted enzymes to maintain high catalytic activity at lower temperatures [81, 75, 76, 83]. Common observed trends include : increased clusters of glycine and a reduced number of proline residues, especially in loop regions to enhance local mobility, a reduced number of charged residues like arginine on the surface, and exposure of bulky hydrophobic residues to reduce packing of the core.

Recently, it has been argued for a funneling picture as an explanatory model of the folding and catalytic function for cold adapted enzymes in connection with calorimetric experimental data [84, 85]. Essentially, D'Amico et al.'s [84] funneling picture for cold enzymes activation is relying on the fact that the psychrophilic protein is accommodated by a larger configurational entropy, a larger ruggedness of the landscape, and a smaller energy gap as compared to mesophilic and thermophilic proteins. At the same time there are local flexibility changes in the vicinity of the active site. The heat liability of both folding and catalytic activity is however such that these functions are more constrained in temperature scale.

A brief review of the approach

In this thesis we study a specific set of homologous protein sequences from mesophilic and psychrophilic organisms that differ by a relatively few substitutions in order to most effectively separate temperature adaptive substitutions from other evolutionary variations. The approach we opt for is to find good observables that can characterize these proteins (enzymes) from differently adapted species. Specifically, we will be interested in identifying localized position dependent excess mean property sequence profile differences and cross species variations between psychrophilic and mesophilic enzymes. In this way, we can measure the differences between the two temperature groups of enzymes in terms of drift of the centroid mean property

sequence and covariance matrix relative to standard mesophilic temperature. We refer to the localized mean deviated property sequence sites as excess mean values (EMVs). These EMVs are due to local and global pressures from evolution without affecting the overall fold of the protein family and evolutionary selection with respect to protein (enzyme) functionality. The observed EMVs can be used to identify where the underlying fluctuations are mainly supported in the 3D structure of a representative protein (from cold/warm) by mapping the EMVs onto sequence positions in the MSA.

We start by presenting a viable way from biological categorical sequence data to numerical values through evolutionary and probability parameterization of these observables. We will be especially interested to model environmental temperature effects, like cold adaptation, into Markov models for standard phylogeny of protein sequences. The complexity of the index space of a protein sequence family makes it very important to have an efficient and simple language to express relations. We will, therefore, represent the amino acid $i(l,s)$ at sequence position s and species l in a MSA, numerically by amino acid unit counts, defined as $\mathbf{Y}_{l,s} = \mathbf{Y}_{i(l,s)}$, where $\mathbf{Y}_{i(l,s)} = (\delta_{i(l,s),i})$. Here $\delta_{i(l,s),i}$ is one if amino acid $i(l,s)$ is i, otherwise zero and $i \in \mathcal{A}$, one of the 20 (or 21 if gap is included) known amino acid from the set of characters $\mathcal{A}$. The unit counts will serve as our basic observable through which other sequence observables, like physico-chemical observables, can be obtained via linear transformations. As we shall see, many "discoveries" can be made simply by presenting the data and the model in a clear way. With the above representation, we have that the average over the observed present time leaf distribution of the protein amino acids at (l,s) is given by the amino acid distribution $\mathcal{E}[\mathbf{Y}_{l,s}] = \mathbf{p}^{l,s}$, where $\mathcal{E}[\cdot]$ denotes expectation with respect to phylogenetic distributions. Here for completeness, we have taken into account that $\mathbf{p}^{l,s}$ will vary both on subset of species and positions. Based on reversible Markov models for protein sequence evolution on a phylogenetic tree [68], we will carry out correlation analysis of a MSA with cold adapted representatives. Note that since we are interested in the EMVs and cross species variations, the pair marginals of the unit counts and related observables are sufficient to describe these variations if time-dependent irreversible changes are introduced as localized events in time.

We will also study how the observed differences between the cold and warm adapted enzymes are due to evolutionary constraints both on the structural phenotype and the genotype. For us this genotype will be given by the protein sequence via the unit count sequence and the phenotype is the state of the protein which in a simple model will be described by the folded contact network. An interesting observable with respect to the genotype will be property hydrophobicity, which has long been considered as one of the primary driving force in the folding proteins. Given a vector $\mathbf{q} \in \mathbb{R}^{20}$ of amino acid properties, we can convert a MSA with unit counts $\mathbf{Y}_{l,s}$ to hydrophobicity sequences through unit count projections as

$\mathbf{h}_l^{(n)} = (\mathbf{Y}_{l,s} \cdot \mathbf{q})_{s\in\{1,\ldots,n\}}$, where the superscript (n) indicates n sequence sites. We will be interested in both cluster dependent EMVs and position dependent common sequence means of sets of aligned protein sequences that share common folds.

We need the position dependent mean hydrophobicity to make possible the analysis of those positions structurally and functionally important in the given fold or protein family through the local exponential parameter profile $\boldsymbol{\beta} = (\beta_s)$ in the distribution. The local exponential probability distribution (with β_s as parameter) will be derived from maximum entropy (MaxEnt) principle for sequence statistics evolutionary theory based on the frustration and mean free energy observables. Such a local exponential distribution in context of phylogenetic inference has been earlier derived by Bastolla and coworkers [60], which is based on the correlation of the mean hydrophobicity profile of protein sequences with the same fold and principal eigenvector (PE) of the contact map $C^f = (C^f_{st} \in \{0,1\})$ of their folded structure (indicated by the superscript f). However, since natural selection may favor only marginally stable proteins, we will also take into account the liquid partially folded conformations. The energy in the folded and liquid states are known to be dominated by a hydrophobicity interaction matrix, like Miyazawa and Jernigan (MJ) matrix [69], contracted on the contact network C. In the simple linear model for the pairwise hydrophobicity interactions this turns out to be possible to decompose into an effective local interaction of the hydrophobicity sequence with the contact density of the network. From the observation of excess mean hydrophobicity at a given sequence position, we will find from our model β_s linearly related to the difference of dynamic mean liquid contact density and the transversal evolutionary mean of the folded contact density (i.e. δc_s). The local exponential parameter β_s contains in principle external environmental dependence due to for example temperature. However, if we want to consider EMVs due to environmental and evolutionary effects, we cannot avoid keeping the effects of the EMVs in say different (may be overlapping of very slightly shifted) clusters in the evolutionary tree.

We will initially apply singular value decomposition (SVD) to cross species covariance matrix and filtering in order to decorrelate and remove the eigensequences buried in evolutionary tree. The average hydrophobicity profile is then computed from the first few significant eigensequences corresponding to the largest eigenvalue of the cross species hydrophobicity covariance matrix. The filtering action is important from sequence analysis point of view. We will use a translational invariant, Gaussian filter. The original idea was to motivate local averaging filters, like wavelets or local fixed cavity fields, by physical local interactions in a way which is akin to what amino acids actually "feel". The introduction of such a local filter will be more true to the actual network interactions, like charge or isoelec-

tric point (Ip) interactions and configurational interactions, especially at the molecular solvent surface, at domain interfaces, at secondary structure boundaries, at active sites, and at binding sites.

A more selfconsistent way of obtaining low rank mean sequence profiles is by use of the factor analysis (FA) [48, 50, 49]. Latent variable models, like FA, have a wide spectrum of application in data analysis. The FA model can be interpreted in our setting by that we are given a model where the mean deviated property sequences, say hydrophobicity can be decomposed into EMVs and a noise term, $\delta\mathbf{h}_s = \delta\boldsymbol{\mu}_s + \boldsymbol{\varepsilon}_s$, where $\boldsymbol{\varepsilon}_s$ is assumed to be white Gaussian noise with covariance Ψ. In an FA model, the EMVs are modelled by $\delta\boldsymbol{\mu}_s = A\mathbf{f}_s$, where the latent vectors $\mathbf{f}_s$ are assumed to have a zero-mean, unit variance Gaussian distribution, i.e. $\mathbf{f}_s \sim \mathcal{N}(0, I)$. Conditional on $\mathbf{f}_s$, the mean deviated property sequences are independently distributed as $\mathcal{N}(A\mathbf{f}_s, \Psi)$. Unconditionally, the $\delta\mathbf{h}_s \sim \mathcal{N}(0, AA^T + \Psi)$, where the cross-species correlations are given by the covariance matrix of the EMVs, $\mathcal{E}[\delta\boldsymbol{\mu}_s\delta\boldsymbol{\mu}_s^T] = AA^T$. The model parameters $\Theta = (A, \Psi)$ can be estimated in a maximum likelihood sense, using for instance the expectation maximization (EM) algorithm [40, 41]. Given, A and Ψ, the expected value of the factors can be computed through the linear projection $\mathcal{E}[\mathbf{f}_s \mid \mathbf{h}_s] = G\delta\mathbf{h}_s$, where $G = A^T(AA^T + \Psi)^{-1}$ are the coefficients in (multivariate) regression of the factors on the sequence variables. These mean profiles and the corresponding factor loadings can be used to classify proteins within a family by identifying significant localized sequence profile differences and cluster drift between psychrophilic and mesophilic species. Note that in the case of a large number of contributions to the construction of the observations and their corresponding estimators both along the sequence and across the clusters in a family of sequences, we will argue that a structured Gaussian model is sufficient for estimation of many gross observables. We will stress that this does not mean that the basic statistical distribution is close to a Gaussian distribution locally, only that a structured cluster based mean deviation property profile is a good description together with structured Gaussian noise across and along the sequences for the observables and estimators of interest. Moreover, it is assumed that the posterior distribution and the corresponding estimators are well described, for the observables, we are concentrating our intention on. It has been earlier found that this point of view is a good description asymptotically in the inverse number of data involved in the observables [71]. In fact, we will find that the structured factor model of clusters also fits well with the asymptotic decomposition of our basic sequence probability distribution model based on the fitness landscape induced by a funneling picture of protein folding.

1.2 Summary of the publications

Paper 1

Said H. Ahmed and Tor Flå, *Estimation of evolutionary average hydrophobicity profile from a family of protein sequences,* PRIB 2007, LNBI 4774, 158–165, Springer-Verlag Berlin Heidelberg 2007.

In this paper, a method to estimate the evolutionary average hydrophobicity profile from set of aligned protein sequences is presented, which is based on SVD and cavity filtering. The idea is to use the eigensequences related to the inter-species hydrophobicity sequence correlation matrix to remove the evolutionary noise from the sequences and hence avoid inspection of a large database to compute the average hydrophobicity. We performed cavity filtering on the average hydrophobicity profile. We motivate the fixed cavity fields by physical local interactions in a way which is akin to what the amino acids actually "feel".

We tested the method on aligned sequences from elastase family. The method effectively removed the evolutionary noise in the hydrophobicity profiles.

Paper 2

Said H. Ahmed and Tor Flå, *Evolutionary parameters in sequence families : Cold adaptation of enzymes,* PRIB 2009, LNBI 5780, 1–12, Springer-Verlag Berlin Heidelberg 2009.

In this paper, correlation analysis of sets of aligned sequences for a protein family is presented, which is based on essentially reversible Markov models for protein sequence evolution on a phylogenetic tree. The goal is to study environmental temperature effects on cold adapted enzymes. In this connection, we present a viable way from categorical biological, sequence data to numerical values through evolutionary and probability parameterization of features and study these features through multiscale and multivariate methods.

Results from this study showed correlations across and along the aligned family of distinct (i.e. elastases and trypsins) protein sequences from differently adapted species. The observed variations was described as centroid drifts in terms of mean and covariance matrix based on physico-chemical properties of the amino acids, indicated by deviations from their constant mean values and clusters of cold and warm adapted enzymes. We found that these variations were mainly determined by a few conserved sites within each cluster of cold and warm adapted enzymes, especially a few amino acid substitutions around the active site positions.

Paper 3

Tor Flå, Said H. Ahmed *Evolution of cold adapted protein sequences,* Sequence and Genome analytics : methods and application II (ed. Gabriel fung, ISBN 9780-9807330-6-8). Concept Press Ltd (2010).

In this paper, a statistical evolutionary model is presented, which describes self-consistently both for transition and equilibrium probabilities of protein sequences, how the notion of a fitness landscape can be introduced for monomorphic populations. The fitness landscape is associated with constraints parameterized through the MaxEnt principle by observables from protein funneling. The concept of funneling on the phenotype state space of the contact network observables conditional on the genotype sequence is developed. This is done in such a way that we have the possibility to define energy and energy variance functionals including both local profile and global means and the possibility to derive different statistical contributions for the property sequences and for the contact network case of both a dynamic and evolutionary nature. From this model, we find a new description of protein phylogenetic statistics given a tree with the possibility of irreversible environmental and effective population changes localized in time. This forms a model based on sequence statistics, phylogenetics and clustering including fitness constraints and environmental changes like temperature.

Paper 4

Said H. Ahmed and Tor Flå, *Position dependent mean hydrophobicity and structural profiles,* unpublished manuscript draft.

In this paper, a method to estimate grand and cluster dependent mean hydrophobicity profiles is presented, which is based on factor analysis (FA) and filtering along the the factor sequences. Unlike SVD and PCA, the FA gives a self-consistent way for cross species noise subtraction in the eigensequences. Furthermore, it is a probabilistic model for a linear statistical problem. In this case, the estimated factor sequence is merely one property, typically the mean or mode of an entire probability distribution. We use a one factor model to extract a mean hydrophobicity profile that is compatible to the structural profile of a given fold and a two factor model to extract cluster (of cold/warm) dependent mean hydrophobicity profiles. In the former case (i.e. the one factor model), we show that the extracted mean hydrophobicity profile is directly related to the difference of mean liquid contact density and the transversal evolutionary mean of the folded contact density through the local exponential distribution parameter profile $\boldsymbol{\beta} = (\beta_s)$. This parameter contains in principle external environmental de-

pendence due to for example temperature. In the latter case (i.e. the two factor model), an orthogonal rotation is performed on the two extracted cluster dependent factor sequences. The goal with the orthogonal rotation is to get a simple structure, that is, to observe changes in the psychrophilic factor sequence relative to standard mesophilic factor sequence.

We demonstrate that these factor sequences and the corresponding factor loadings can be used to to classify members of a protein family by identifying significant localized hydrophobicity profile differences and cluster drift between psychrophilic and mesophilic species.

Chapter 2

Paper 1
Estimation of Evolutionary Average Hydrophobicity Profile from a Family of Protein Sequences

Chapter 3

Paper 2
Evolutionary Parameters in Sequence Families : Cold adaptation of enzymes

Chapter 4

Paper 3
Evolution of cold-adapted protein sequences

Chapter 5

Paper 4
Position dependent mean hydrophobicity and structural profiles

Chapter 6

Further work and perspectives

In this work, we have not studied the clustering effects as localized to different domains. If, as often is the case, the protein consists of two relatively recent fused domains each with its own evolutionary history and tree, it will make sense to divide the clustering analysis according to the protein domain boundaries and let the sequence statistics as a first model consists of two independent parts with its own clusters. If the fusion of the domain is not recent, a consensus tree and clustering might again be a useful model although one might have to be careful and interpret clustering effects on different domains.

It is possible to generalize the funneling model to more complicated parameterized energy models than the linear hydrophobicity model. More complicated state descriptions dynamic and heterogenous in time and sequence position and even with several subclusters can also be formulated. This will be needed for example to compare with more realistic MD simulations and proteins with many domains and binding/active sites.

Close links between different statistical techniques for modeling multidimensional data sets, like FA, probabilistic PCA and independent component analysis (ICA), should be examined. In particular, it should be investigated whether tools known from different techniques could be combined into more powerful machine learning algorithms.

Bibliography

[1] Kimura, M. The neutral theory of molecular evolution. Cambridge, Cambridge University Press, 1983.

[2] Anfinsen, C. B. Principles that Govern the folding of protein chains. Science, 181, 223–230, 1973.

[3] Onuchic, J. N, Wolynes P. G. Theory of protein folding. Curr. Opin. Struc. Biol. 14, 70–75, 2004.

[4] Luthey-Schulten, Z., Wolynes, P. G. Theory of protein folding - the energy landscape perspective. Annu. Rev. Phys. Chem., 48, 545–600, 1997.

[5] Koretke, K. K., Luthey-Schulten, Z., Wolynes, P. G. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. Proc. Natl. Sci. USA, 95, 2932–2937, 1998.

[6] Wolynes, Peter G. Entropy crises in glass and random heteropolymers. J. Res. Natl. Inst. Stand. Technol., 102, 001, 1997.

[7] Plotkin, S. S., Wang, J., Wolynes, P. G. Statistical mechanics of a correlated energy landscape model for protein folding funnels. J. Chem. Phys., 106, 2932–2948, 1997.

[8] Li, H., Helling, R., Tang, C., Wingreen, N.S. Nature of driving force for protein folding : A result from analyzing the statistical potential. Phys. Rev. Lett. 79, 765–768, 1997.

[9] Saven, J. G., Wang, W. Designing gene libraries from protein profiles for combinatorial protein experiments. Nucleic Acids research, 21, 120–128, 2002.

[10] Rose, G., Geselowitz, A., Lesser, G., Lee, R. and Zehfus, M. Hydrophobicity of amino acid residues in globular proteins. Science 229, 834–838, 1985.

[11] Flå, T., Ahmed, S. H. Evolution of cold-adapted sequences. Sequence and Genome analytics : methods and application II (ed. Gabriel fung, ISBN 9780-9807330-6-8), Concept Press Ltd, 2011.

[12] Overington, J., Johnson, M. S., Sali, A., Blundell, T. L. Tertiary structural constraints on protein evolutionary diversity - templates, key residues and structure prediction. Proc. R. Soc. Lond. Ser. B 241, 132–145, 1990.

[13] Govindarajan, S., Goldstein, R. A. Searching for foldable protein structures using optimized energy functions. Biopolymers, 36, 43–51, 1995.

[14] Goldstein, R. A., Luthey-Schulten, R. A., Wolynes, P. G. Optimal protein-folding codes from spin-glass theory. Proc. Natl. Acad. Sci. USA, 89, 4918–4922, 1992.

[15] Gutin, A., Sali, A., Abkevich, V., Karplus, M., Shakhnovich, E. I. Temperature dependence of the folding rate in a simple protein model : search for a "glass" transition. J. Chem. Phys. 108, 6466–6483, 1998.

[16] Finkelstein, A. V., Gutun, A. M., Badretdinov, A. Y. Why are the same protein folds used to perform functions?. Federation of Eur. Biochem. soc., 1,2, 23–28, 1993.

[17] Shakhnovich, E. I., Gutin, A. M. Formation of unique structure in polypeptide chains: Theoretical investigation with the aid with the aid of a replica approach. Biophys. Chem., 34, 187–199, 1989.

[18] Mirny, L. A., Shakhnovich, E. I. How to derive a protein folding potential? A new approach to an old problem. J. Mol. Biol., 264, 1164–1179, 1996.

[19] Bryngelson, J. D., Wolynes, P.G. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. USA., 84, 7524-7528, 1987.

[20] Taverna, D., Goldstein, R. A. Why are proteins marginally stable? Proteins, 46, 105–109, 2002.

[21] Goldstein, R. A. Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles : Insights from the quasi-chemical approximation. Protein science, 16, 1887–1895, 2007.

[22] Gutell, R., Power, A., Hertz, G.Z., Putz, E. J., Stormo, G. D. : Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Research 20, 5785–5795, 1992.

[23] Berg, O. G., Von Hippel, P. H. Selection of DNA binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol., 193, 723–750, 1987.

[24] Luthy, R., Bowie, J.U., Eisenberg, D. Assessments of protein models with three-dimensional profiles. Nature, 356, 83–85, 1992.

[25] Rost, B., Sander, C. Progress of 1D protein-structure prediction at last. Proteins, 23, 295–300, 1995.

[26] Kyte, J., Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 157, 105–132, 1982.

[27] Eisenberg, D., Wilmanns, M. 3-dimensional profiles from residue-pair preferences - identification of sequences with beta/alpha-barrel fold. Proc. Natl. Acad. Sci. USA, 90, 1379–1383, 1993.

[28] Eisenberg D., Weiss R., Terwillinger T. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad Sci, USA, 81 (1), 140–144, 1984.

[29] Stormo, G. D., Fields, D. S. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem., Sci, 23, 109–113, 1998.

[30] Derrida, B. : Random energy model - an exactly solvable model of disordered systems, Phys. Rev. B, 24, 2613–2626, 1981.

[31] Ptitsyn, O.B. Physical principles of proein sructures and protein folding. J. biosci. 8, 1–13, 1985.

[32] Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. 15, 563–577, 1999.

[33] Casari, G., Sander, C. Valencia, A. A method to predict functional residues in proteins. Nat. Struct. Biol., 2, 171–178, 1995.

[34] Andrade MA, Casari, G., Sander, C., Valencia, A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol. Cybern, 76, 441–450, 1997.

[35] Pei, J., Cai, W., Kinch, L. N., Grishin, N. V. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics, 22, 164–171, 2006.

[36] Donald, J. E., Shakhnovich, E. I. Predicting specificity-determining residues in two large eukaryotic transcription factor families. Nucleic Acids Res, 33, 4455–4465, 2005.

[37] Nadaraya, E. A. On estimating regression, "Theory of probability and its applications", 10, 186–190, 1964.

[38] Watson, G. S. Smooth regression analysis, Sankhya series A, 26, 101–116, 1964.

[39] Johnson, R.A., and Wichern, D.W. Applied Multivariate Statistical Analysis. Fifth ed., Prentice Hall, Upper Saddler River, New Jersey, 2002.

[40] Dempster, A. P., Laird, N. M., Rubin, D. B. Maximum likelihood from incomplete data via EM algorithm. Journal of the Royal Statistical Society series B, 39, 1–38, 1977.

[41] Ghahramani, Z., Hinton, G. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96, Dept. of computer science, University of Toronto, 1996.

[42] Blake, A., Isard, M. Active contours, the application from graphics, vision, control theory, and statistics to visual tracking of shapes in motion. Springer-Verlag London, 1998.

[43] Shannon, E., Weaver, W. The mathematical theory of communication. University of Illinois Press, Urbana, 1949.

[44] Inge Helland. Partial least squares regression and statistical models. Artikkel i "Scabdinavian Journal of Statistics", 1990.

[45] Guttman, L. Multiple rectilinear prediction and the resolution into components. Pshychometrika, 5, 75–99, 1940.

[46] Teichert F. and Porto M. Vectorial representation of single-and multi-domain protein folds. Eur. Phys. J.B 54, 131–136, 2006.

[47] Jöreskog, K. G. Some contributions to maximum likelihood factor analysis. Psychometrika, 32, 443–482, 1967.

[48] Jöreskog, K. G., Lawley, D. N. New methods in maximum likelihood factor analysis. British Journal of mathematics and statistical psychology, 21, 85–96, 1968.

[49] Lawley, D. N., Maxwell, A. E. Factor analysis as a statistical method. 2nd ed. London, Butterworths, 1971.

[50] Jöreskog, K. G. Factor analysis by least sequares and maximum likelihood. In *statistical methods for digital computers*, edited by Enslein, K., Ralston, A. and Wilf, H. S., New York: John Wiley, 1975.

[51] Bartlett, M.S. The statistical conception of Mental Factors. British Journal of Psychology, 28, 97–104, 1937.

[52] Bartlett, M.S. A note on multiplying factors for various Chi-squared approximations. Journal of the Royal statistical society, 16, 296–298, 1954.

[53] Akaike, H. Factor analysis and AIC. Psychometrika, 52, 317–332, 1987.

[54] Strang, G., Nguyen, T. Wavelets and Filter Banks. Wellesley-Cambridge Press, 1997.

[55] Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell., 11 (7), 674–693, 1989.

[56] Opper, M., Winther, O From naive mean field theory of the TAP equations. MIT press, Cambridge, Massachusetts London, England, 2002.

[57] Scharf, Louis L. Statistical signal processing, Detection, Estimation, and Time series analysis. Addison-Wesley Reading, MA, 1991.

[58] Peebles, P. Z. Probability, random variables and random signal principles. 3rd edition, McGraw-Hill international editions, 1993.

[59] Durbin, R., Eddy, S., Krogh, A., Mitchison, G. Biological sequence analysis, " Probabilistic models of proteins and nucleic acids", Cambridge university press, 1998.

[60] Bastolla, U., Porto, M. Roman, HE., Vendruscolo, M. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. Evol. Biol. 6, 43, 2006.

[61] Parisi, G., Echave, J. Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. 18, 750–756, 2001.

[62] Dokholyan, N.V., Shakhnovich, E. I. Understanding hierarchical protein evolution from first principles. J. Mol. Biol. 321, 2001.

[63] Overington, J., Johnson, M. S., Sali, A., Blundell, T. L. Tertiary structural constraints on protein evolutionary diversity - templates, key residues, and structure prediction. Proc. R. Soc. Lond. Ser. B, 241, 132–145, 1990.

[64] Branden, C., Tooze, J. Introduction to protein structure, Garland pub. inc., 1999.

[65] Lio, P., Goldman, N. Models of molecular evolution and phylogeny. Genome Res. 8, 1233–1244, 1998.

[66] Guttman, L. Best possible systematic estimates of communalities. Psychometrika, 21, 272–278, 1956.

[67] Koshi, J.M., Goldstein, R. A. Models of natural mutation including site heterogeneity. Proteins, 32, 289–295, 1998.

[68] Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17, 368–376, 1981.

[69] Miyazawa, S., Jernigan, R.L. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, 18, 534–552, 1985.

[70] Fauchere, J.L., Pliska, V. Hydrophobic parameters of amino acid side chain from the partitioning N-acetyl amino acid amides. Eur. J. Med. Chem., 18, 369–375, 1983.

[71] Fossgaard, E., Flå, T. An invariant bayesian model selection principle for Gaussian data in a sparse representation. IEEE Transactions on Information Theory, 52 (8), 3438–3455, 2006.

[72] Ahmed, S.H., Flå, T. Evlutionary parameters in sequence families : Cold-adapted enzymes. PRIB 2009, LNBI 5780, 1–12, 2009.

[73] Jaenicke, R., Bohm, G. The stability of proteins in extreme environments. Curr. Opin. Struct. Biol. 8, 738–748, 1998.

[74] Lonhienne, T, Gerday, C, Feller, G. Psychrophilic enzymes: revisiting the thermodynamic parameters of activation may explain local flexibility. Biochim. Biophys. Acta, 1543, 1–10, 2000.

[75] Georlette, D., Blaise, V., Collins, T., D'Amico, S., Gratia, E., Hoyoux, A., Marx, J.C., Sonan, G., Feller, G., Gerday, C. Some like it cold: biocatalysis at low temperatures. FEMS icrobiol. Rev., 28, 25–52, 2004.

[76] Feller, G., Gerday, C. Psychrophilic enzymes: molecular basis of cold-adaptation. Cell Mol. Life Sci, 53, 830–841, 1997.

[77] Ahmed, S. H., Flå, T. Estimation of evolutionary average hydrophobicity profile from a family of protein sequences, PRIB 2007, LNBI 4774, 158–165, 2007, Springer-Verlag Berlin Heidelberg, 2007.

[78] Siddiqui, K. S., Feller, G., D'Amico, S., Gerday, C., Giaquinto, L., Cavicchioli, R. : The active site is the least stable structure in the unfolding pathway of a multidomain cold-adapted α-Amylase. J. Bacteriology, 187 (17), 6197–6205, 2005.

[79] Gerday, C., Aittaleb, M., Bentahir, M., Chessa, J.P., Claverie, P., Collins, T., D'Amico, S., Chessa, J.P., Dumont, J., Garsoux, G., Georlette, D., Hoyoux, A., Lonhienne, T., Meuwis, M.A., and Feller, G. Cold-adapted enzymes: from fundamentals to biotechnology. Trende Biotechnol., 18, 103–107, 2000.

[80] Olufsen, M., Smalås, A. O., Moe, E., Bransdal, B. O. Increased flexibility as a strategy for cold adaptation. J. Biol. Chem, 280, 18042–18048, 2005.

[81] Somero, G. N. Proteins and temperature. Annu, Rev. Physiol. 57, 43–68, 1995.

[82] Georlette, D., Damien, B., Blaise, V., Depiereux, E., Uversky, V., Gerday, C., Feller, G. Structural and functional adaptations to extreme temperatures in psychrophilic, mesophilic and thermophilic DNA ligases. J. Biol. Chem. 278, 37015–37023, 2003.

[83] Siddiqui, K.S., Cavicchioli, R. Cold-adapted enzymes. Ann. Rev. Biochem., 75, 403–433, 2006.

[84] D'Amico, S., Collins, T., Marx, J., Feller, G., Gerday, C. Activity-stability relationships in extremophilic enzymes. J. Biol. Chem., 278, 2891–7896, 2003.

[85] Feller, G. Life at low temperatures : is disorder the driving force? Extremophiles, 11 (2), 211–216, 2007.

[86] Schrøder, H-K., Willassen, N. P., Smalås, A. O. Residue determinants and sequence analysis of cold-adapted trypsins. Extremophiles 2, 05–219, 1999.

[87] Leiros, H. K. S., Mcsweeney, S. M., Smalås, A. O. Atomic resolution of trypsin provide insight into structural radiation damage. Acta Crys. SectD 57, 488, 2001.

[88] Beitz, E. TEXshade : shading and labeling of multiple sequence alignments using Latex2ϵ. Bioinformatics, 16, 135–139, 2002.

[89] Delano, W. L. The PyMOL User's Manual. DeLano Scientific, San Carlos, CA., 2002.

ISBN 978-82-8236-036-4