HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTER THESIS

Mr.
Silvio Oswald

Investigations on quantilized energetically sub-profiles in Pfam families and discussion concerning their information content

Mittweida, 2014

Department of Mathematics, Natural and Computer sciences

MASTER THESIS

Investigations on quantilized energetically sub-profiles in Pfam families and discussion concerning their information content

Author:
Mr. B.Sc.

Silvio Oswald

Course of studies:
Molecular Biology/Bioinformatics

Seminar group:
MO12-w1m

First examiner:
Prof. Dr. rer. nat. Dirk Labudde

Second examiner:
M.Sc. Florian Heinke

Submission:
Mittweida, 25.08.2014

Defence/Evaluation:
Mittweida, 2014

If the doors
Of perception were cleansed
Everything would appear
To man as it is:
Infinite

The Marriage of Heaven and Hell

William Blake

Bibliographic Description:

Oswald, Silvio:
Investigations on quantilized energetically sub-profiles in Pfam families and discussion concerning their information content. - 2014. - 59 pages of content, 3 pages of directories

Hochschule Mittweida (FH), Department of Mathematics, Natural and Computer sciences,

Master Thesis, 2014

Abstract

Protein structures are essential elements in every biological system evolved on earth, where they function as stabilizing elements, signal transducers or replication machineries. They are consisting of linear-bonded amino acids, which determine the three-dimensional structure of the protein, whereas the structure in turn determines the function. The native and biological active structure of a protein can be understood as the folding state of a polypeptide chain at the global minimum of free energy.
By means of *protein energy profiling*, which is an approach derived from statistical physics it is possible to assign a so called energy profile to a protein structure. Such an energy profile describes the local energetic interaction features of every amino acid within the structure and introduces an energetic point of view, instead of a structural or sequential onto proteins.
This work aims to give a perspective to the question of how we may gain pattern information out of energy profiles. The concrete subjects are energy-mapped Pfam family alignments and investigations on finding motifs or patterns in discretizised energy profile segments.

Zusammenfassung

Proteinstrukturen sind zentrale Bausteine eines jeden biologischen Systems, welches sich auf der Erde entwickelt hat, wo sie als stabilisierende Elemente, Signaltransduktoren oder Replikationsschnittstellen fungieren. Sie sind aus linear verknüpften Aminosäuren aufgebaut, welche die Struktur eines Proteins mitbestimmen, wobei die Struktur wiederrum die Funktion eines Proteins bestimmt. Die native und biologisch aktive Struktur eines Proteins kann als der Faltungszustand verstanden werden, den eine Polypeptidkette am globalen Minimum der freien Energie einnimmt.

Mit der Methode des *protein energy profiling* – einem Ansatz aus der statistischen Physik – ist es möglich jeder aufgeklärten Proteinstruktur ein sogenanntes Energieprofil zuzuordnen. Solch ein Profil beschreibt die lokalen energetischen Wechselwirkungen, die Aminosäuren in der Struktur unterliegen und stellt neben der sequenziellen und strukturellen Sichtweise auf Proteine eine weitere Abstraktionsebene dar.

Diese Arbeit zielt darauf ab die Frage, wie sich Motivinformation aus Energieprofilen extrahieren lässt, zu konkretisieren. Das Hauptaugenmerk liegt dabei auf Pfam Familienalignments, denen ihre korrespondierenden Energieprofilsegmente zugeordnet wurden; sowie auf Untersuchungen zur Mustersuche in diskretisierten Energieprofilsegmenten.

Acknowledgments

First of all I would like to own gratitude to Professor Dirk Labudde at the University of Applied Sciences Mittweida for giving me the opportunity to work on this sensitive, but challenging and interesting subject. His ability to calm down students, who were diving to deep in the spheres of scientific methodology and his way to take a closer look at the pure essence of biology often, impressed me. No lesser gratitude I own to my tutor Florian Heinke, who had constantly monitored and inspired this master thesis with indispensable ideas.

I would also like to thank my fellow students Alexander Hampel, Mathias Langer and Sebastian Bittrich for discussions, guidance and motivation throughout the time of working and writing this thesis. Especially Mr. Bittrich shall be plait with wreaths for proofreading these pages.

Contents

List of Figures

List of Tables

1 Motivation

Since life-forms, as we know have been evolved from simple self-organizing protocells proteins are indispensable components of biological systems, where they function as stabilizing elements, signal transducers or replication machineries. Even considering, that the discovery of biological functionality of protein structures is an ongoing and probably the most challenging issue in the wide field of molecular and structure biology, a deep reflection of processes like protein folding or catalytic effects is still to be found.

The proclaimed aim of structure-based drug design is to design a functional protein towards a user-defined goal. In this process accurate knowledge about the native conformation of a target protein is fundamental. Despite the efforts and progress made over the last decades the protein folding problem remains unsolved, which indicates that de-novo structure prediction is one of the most valuable tasks in modern biology. Computational structure modeling and experimental structure determination have proven themselves as indispensable approaches for making up this lack of information. However, the applicability of in-silico methods is limited and dependent on the quantity and quality of underlying template data. Experimental approaches are time demanding, costly and the probability of success is strongly influenced by the multilayered molecular nature of the protein. Thus, a lot of research is focused on understanding the sequence encrypted features which determine fold and function of a protein.

Databases like the Pfam- [1] and the CATH-database [2] are providing distinctions of protein structures into families of folds based on both sequential and structural point of view. Since Pfam was released in 1997 the database contains 14.831 manually curated protein families as of release 27.0. [3] The classification of a protein sequence into a specific Pfam-family is based on multiple sequence alignments. Another approach to distinguish protein structures is provided by the CATH-database where structures are considered as realizations of a widely ranged folding space. The CATH Protein Structure Classification is a semi-automatic, hierarchical classification of protein domains published in 1997 and associates every structure with a unique CATH-Identification code representing the Class, Architecture, Topology and Homology. CATH shares many broad features with its principal rival, SCOP, but there are also many areas in which the detailed classification differs greatly. Further deliberations on the distinction approaches of these databases are given in a later chapter.

Even though the Pfam- and CATH-approaches are well-established and proven methods, it is possible to approach protein structures in different ways than based on sequence alignments or the classification into fold-classes. In contribute to this research there is another approach called energy profiling to describe protein structures in an energetic way. [4] One main-goal of this work is the application of the energy-profiling approach on protein families and provided by the Pfam database.

With the description of protein structures by means of energy profiling it is possible to assign a one-dimensional vector of pseudo-energies to a three-dimensional structure. Although the energy-approach is at the moment only applicable to globular proteins; the aim is the identification of protein family-specific energy profile fingerprints that are masked by sequence diversity and remain undetected in classic sequence alignments. Based on these identified fingerprints, it shall be allowed to aim at reinvestigating amino acid mutagenity and the important role of sequence diversity with respect to stabilizing energetic conservations and variations in protein families. Insights derived from these investigations could aid in decoding the structure and function determining information encrypted in a protein’s sequence.

2 Fundamentals and Methodology

For laying the foundations towards a better understanding of the further work it is necessary to take a closer look on the underlying terminology. That includes the theoretical understanding of protein structures, their experimental determination and their folding processes, energy profiling and the discretization of energy profiles, the data resources, which are the Pfam- and the CATH-database and a precisely look at techniques like sequence logos [5] and position weight matrices [6].

2.1 Protein Structures and Folding

Proteins were first described by the Dutch chemist Gerardus Johannes Mulder and named by the Swedish chemist Jöns Jacob Berzelius in 1838. Mulder carried out elemental analysis of common proteins and found that nearly all proteins had the same empirical formula, $C_{400}H_{620}N_{100}O_{120}P_1S_1$. He came to the erroneous conclusion that they might be composed of a single type of molecule.

The term "protein" to describe these molecules was proposed by Berzelius, who derived the term from the greek word πρώτειος (*proteios*), meaning "primary".

Nowadays it is known, that proteins are a three-dimensional composition of the different secondary structure elements helices, strands and coils, where helices can be described as large corkcrew-shaped or cylindrical formations, strands as nearly plain series of amino acids and coils as approximately random conformations of a protein’s structure. In terms of simplification the term amino acid will be replaced by residues. Connected through the peptide bond, which is a rotable covalent connection between two residues, a series of residues is able to fold itself into a native structure shortly after the process of protein biosynthesis.

If we consider a poly-peptide chain with a length of 100 residues and suppose, that every residue is able to move into one of the three states helix, strand or coil, there are 3^{100} or 10^{48} combinatorial possibilities to order the chain in a structural way. The knowledge, that the rotation around a peptide bond occurs with a frequency of 10^{14}/s leads us to the conclusion that random folding will take nearly 10^{26} years to end in a stable and biologically active structure. By adding the fact, that the age of the whole universe is estimated with 13.75 billion or $10^{10.138}$ years it is obvious, that protein folding cannot be described as a random process.

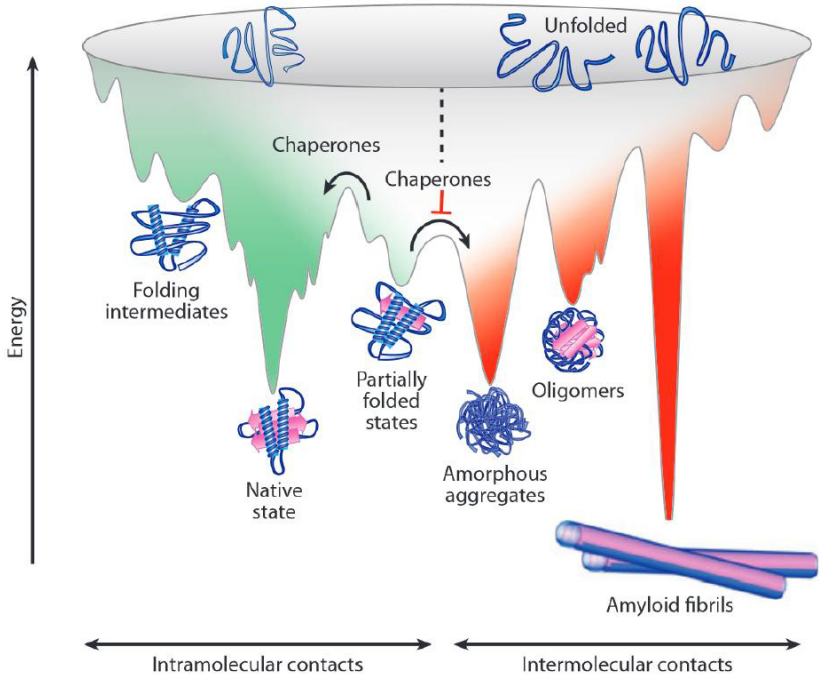


Figure 2.1: Conformation space and folding funnel [7]

Shortly after the process of protein biosynthesis new proteins are folding themselves into their native state. In some cases chaperons are assisting this process to avoid misfolded aggregations. The principle that a stable native structure only exists in an energetic minimum is an important fact to model and design protein structures computationally.

This short thought experiment known as the Levinthal paradox was formulated first by the American molecular biologist Cyrus Levinthal in 1969. [8] His conclusion was that the process of protein folding is not only enforced by random fluctuations in the conformation space, rather it can be described as a path through an energetic landscape (see Figure 2.1). This energetic landscape with its local minima and its global minimum, which represents the native state of a protein structure is called folding funnel.

A mainly enforcing component in protein folding is the hydrophobic collapse, where regions consisting of residues with a hydrophobic physico-chemical character are turning into the inside of the protein and other regions consisting of hydrophilic residues are turning towards the surrounding water solvent. However, the combinatorial explosion given by Levinthal is one problem-to-solve in protein structure prediction; nowadays algorithms in this field are much more sophisticated than early approaches like the calculation of Chou-Fasman [9] preferences or the GOR-method [10] to predict secondary structure elements out of a protein sequence.

Besides a vast number of approaches to predict a proteins tertiary structure with computationally methods like I-TASSER [11], HHpred [12] or SWISS-MODEL [13], only to name a few, modern biology is substantially driven by the experimental determination of structures. Furthermore, nearly all prediction methods are based on experimental data from solved 3D structures. Since the late 1950s, beginning with the structure of sperm whale myoglobin by Sir John Cowdery Kendrew [14] a tremendous amount of protein structures (around 90%) available in the Protein Data Bank [15] have been determined by X-ray crystallography. For establishing this technology in the field of biology Kendrew and Perutz received the Nobel Prize in Chemistry in 1962. In an X-ray diffraction measurement a crystallized protein is mounted on a goniometer and gradually rotated while being bombarded with X-rays. This process produces a diffraction pattern of regularly spaced spots:

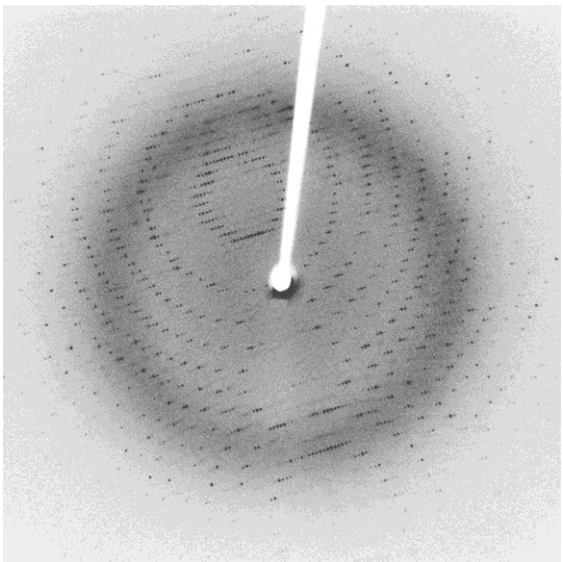


Figure 2.2: X-ray diffraction pattern of a protein crystal [16]

The pattern of spots (reflections) and the relative strength of each spot (intensities) have a significant influence on the later resolution of the structure. Poor resolutions (fuzziness) or even errors may result if the crystals are too small, or not uniform enough in their internal makeup.

These two-dimensional diffraction pattern images taken at different rotations are converted into a three-dimensional model of the density of electrons within the crystal by means of Fourier transform.

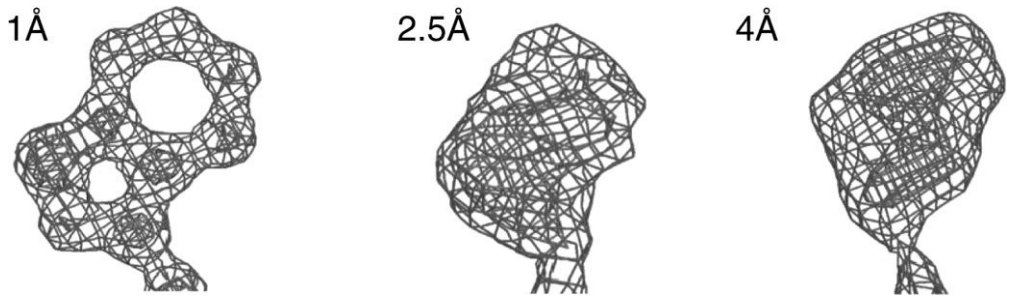


Figure 2.3: Electron densities of tryptophane [17]

The resolution of the electron density measured in Angström is an essential measurement of quality to X-ray solved protein structures determining their later applicability on different tasks.

By using the knowledge of the chemical structure of the 20 canonical amino acids the grid structure of each residue is fitted into the electron density model to get the final structure.

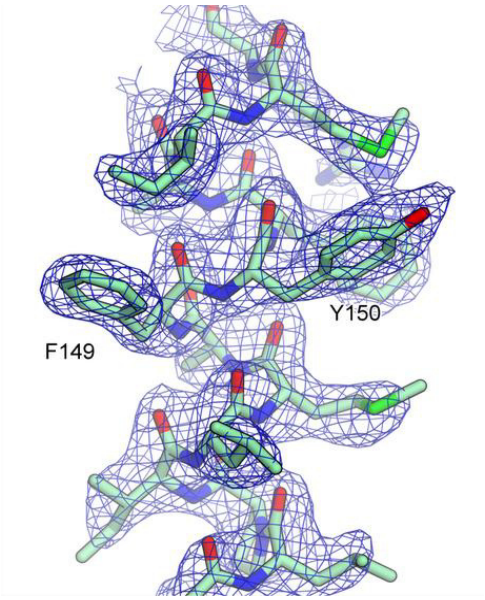


Figure 2.4: Electron density map of a transmembrane helix [18]

The fitting of the residue atomic models into the electron density map is an important step in determining the 3D structure of a protein. Regions with oxygen-atoms (red) have particular high electron densities.

Another way to determine protein structures is the nuclear magnetic resonance spectroscopy (NMR), where the quantum mechanical properties of the central core ("nucleus") of an atom are involved. These properties depend on the local molecular environment, and their measurement provides a map of how the atoms are linked chemically, how close they are in space, and how rapidly they move with respect to each other. Nuclei with an odd number of protons or neutrons have a nonzero nuclear spin. Due to application of a magnetic field, the spins can be aligned and nuclei are assigned with two distinct energetic states: one favorable and one disadvantageous of higher energy opposing the external field. A suitable radio frequency pulse can force the transition between the two states. This phenomenon of nuclear magnetic resonance can be exploited to obtain information about the environment certain nuclei are located in - they are slightly shifted away from the values expected. These chemical shift data is finally used for model building.

Historically, the structures determined by NMR have been, in general, of lower quality than those determined by X-ray diffraction. Conclusively, the PDB today contains about 100.000 solved structures [14], which includes proteins, DNA, RNA and their complexes and is, therefore a valuable source for structure-based research.

2.2 Energy Profiling

This chapter is dedicated to energy profiles as a way to describe protein structures in a more physical manner. The point of departure is the design of force fields, which should lead to the core of the energy profile approach. Finally the discretization of energy profiles as a fundamental topic regarding this work will be explained.

2.2.1 Knowledge-based Force Fields

To create a shared space of terms it is necessary to clarify some basic terms: The term *energy* refers to the conformational energy of an individual polypeptide chain and its interaction energy with the surrounding molecular environment. The energy is a function of the conformational variables of the system like Cartesian coordinates, distances between atoms, binding angles, etc. Taking the derivative of the energy with respect to the conformational variables we obtain the *force field* of the molecule. Energy functions and force fields therefore, are closely related physical quantities which constitute an energetic model of a real physical system. [19]

The modeling of molecular force fields can be approached from two different points of views. The inductive strategy uses the results obtained from quantum-mechanical calculations on small molecules and thermodynamic or spectroscopic data derived from small systems. The resulting data is extrapolated to the macromolecular level under consideration of the hypothesis that the complex behavior of macromolecular systems results from the composition of a vast number of the same type of interactions as found in the most basic molecular systems. [19] The force fields obtained in this approach are called *semi-empirical* force fields. On the other hand, the deductive or knowledge based strategy is resting on the fact, that the forces encountered in large molecular systems are very complicated and to take full account of their complexity the known macromolecular structures are taken as the only reliable source of information. The goal is to extract the forces and potentials stabilizing native folds in solution from a set of known structures.

Although the research based on the inductive approach has led into the development of several semi-empirical force fields, only recently the deductive or knowledge-based approach became a field of intense study due to the growing number of experimentally determined 3D structures available on one hand, and the application of powerful concepts in statistical physics on the other. [19] The energy profiles used in this work belong to knowledge-based approach on force fields.

Based on the principle, that, in equilibrium, thermodynamic systems attain the global minimum of free energy, an adaption to protein solvent system, the so called folding postulate can be described: In equilibrium the native state of the protein solvent system corresponds to the global minimum of free energy. [19]

At the global minimum of free energy the individual molecules, that build the molecular system, may adopt one particular or many different states or conformations. Under physiological conditions soluble globular proteins usually adopt one or several related conformations. In the case of short peptide chains however, the individual molecules are often distributed over a range of dissimilar conformations. [20] This distribution can be described by Boltzmann's principle, an important principle that merges the energy *E* of a system and its probability density function *p.* Using discrete variables Boltzmann's principle can be written as:

$$p_{ijk} = Z^{-1} \exp\left(-\frac{E_{ijk}}{k_B T}\right) \qquad (1)$$

where k_B is Boltzmann's constant, *T* the absolute temperature and the subscripts *i,j,k* corresponds to the variables of the system. The quantity *Z* is called partition function:

$$Z = \sum_{ijk} \exp\left(-\frac{E_{ijk}}{k_B T}\right) \qquad (2)$$

The deductive approach is based on the inverse Boltzmann principle:

$$E_{ijk} = -k_B T ln\left(f_{ijk}\right) + kTlnZ \qquad (3)$$

Here, the energy function E_{ijk} is called potential of mean force [19]. The energy of a state labeled by *i, j, k* is derived from the relative frequencies f_{ijk} obtained from measurements on this state. These frequencies are equivalent to the probability densities p_{ijk} from equation (1), to the effect that, in the limit of infinitely many observations, relative frequencies converges to a probability. The partition function *Z* from equation (2) cannot be obtained from experimental measurements directly.

However, at constant temperature, Z is a constant and does not affect the energy difference between particular states. [20] Choosing $Z = 1$ equation (3) can be simplified to:

$$E_{ijk} = -kTln(f_{ijk}) \qquad (4)$$

Considering the assumption, that the relative state frequencies f_{ijk} shall be equivalent to the probability density function in equation (1) we have to ensure, that the data basis for the application of an appropriate energy calculation is as solid as possible.

However it's impossible to collect an infinite amount of experimental data of the system of interest, it remains as an enormous task to collect a sufficient amount of data. Nevertheless, the collection of protein structures solved over the last decades is a stable foundation for the application of Boltzmann's inverse principle.

2.2.2 Derivation of Energy Profiles

Following Anfinsens statement, that the native protein conformation is determined by the sum of residue interactions many coarse- and fine-grained models to describe residue interactions were developed and published. [21] They are based either on first principles approaches using physical laws or make use of knowledge of existing experimentally derived structures and statistical analysis.

In the sub-chapter before the inverse Boltzmann principle was derived from a general form of Boltzmann's principle. Now the relative state frequencies f_{ijk} from equation (4) will be applied specifically to protein structures in a solvent system. An inside/outside distinction for residues has been applied as the definition of the state of a residue.

Based on [22, 23] such an inside/outside property for generating the residue buriedness distributions and, hence, following the pseudoenergy approximation is defined:

$$f(i) = \begin{cases} n_{inside\,,i} + +, & ||C_{\propto,i} - c|| < 5\,\text{Å} \vee \left(C_{\propto,i} - C_{\beta,i}\right)\left(C_{\propto,i} - c\right) < 0, \\ n_{outside\,,i} + +, & else, \end{cases} \qquad (5)$$

Let *i* denote one of the 20 canonical residues. $n_{inside,i}$ and $n_{outside,i}$ describe the absolute frequency of residue *i* being assigned as “inside” or “outside” by the inside/outside property. Finally c denotes the center of mass of all C_α atoms within a 5 Å sphere surrounding *i*. Applying the inverse Boltzmann principle (equation 4), the pseudoenergy e_i of *i* can be approximated as follows:

$$e_i = -k_B T ln\left(\frac{n_{inside\ ,i}}{n_{outside\ ,i}}\right) \tag{6}$$

Since k_B and *T* are declared as constants in this model, both can be omitted from the calculation:

$$e_i^* = -ln\left(\frac{n_{inside\ ,i}}{n_{outside\ ,i}}\right) \tag{7}$$

The energy of the pairwise interactions of *i* to other residues corresponds to the environment of *i a*nd the environments composition inside the structure. [19] Thus, the expected tendency value *P* of the observed environment composition correlates with the interaction energy of *i*. *P* can be approximated by the derived amino acid distributions [24]:

$$P_{k\epsilon Env} = \prod_{k\epsilon Env} p_k = \prod_{k\epsilon Env}\left(\frac{n_{inside\ ,i}}{n_{outside\ ,i}}\right)$$

$$lnP_{k\epsilon Env} = \sum_{k\epsilon Env} ln\left(\frac{n_{inside\ ,i}}{n_{outside\ ,i}}\right). \tag{8}$$

In analogy to equation (7) the energy of the environment can be derived from equation (4)

$$E_{Env} = -lnP_{k\epsilon Env} \tag{9}$$

and hence, can be written as:

$$E_i = -|Env|ln\left(\frac{n_{inside\ ,i}}{n_{outside\ ,i}}\right) - \sum_{k\epsilon Env} ln\left(\frac{n_{inside\ ,i}}{n_{outside\ ,i}}\right). \tag{10}$$

The last task-to-solve towards an applicable energy function is to define the environment of a residue. Therefore a contact function $g(i,j)$ adapted from [25] is applied:

$$g(i,j) = \begin{cases} 1, & \left\| C_{\beta,i} - C_{\beta,j} \right\|_{D_E} \leq 8\,\text{Å} \\ 0, & else \end{cases}. \qquad (11)$$

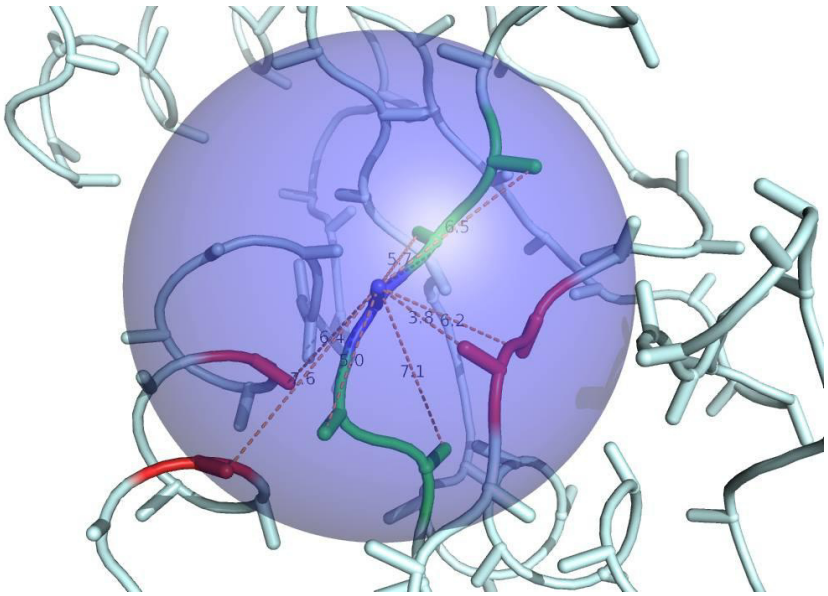


Figure 2.5: 8 Å sphere from the contact function [26]

Every residue *j*, which is within an approximated C_β -distance-measured 8 Å sphere around a residue *i* is taken into the environmental account of *i*. Sequentially close (green) as well as spatially close residues from other parts of the protein (red) are found within the 8 Å sphere.

In contrast to the inside/outside property here a sphere of 8 Å is chosen to decide, wether residue *i* and *j* have a contact or not. This contact measurement uses approximated distances between the C_β atoms of the residues. Finally the total energy E_i^* of a residue *i* is given by:

$$E_i^* = \sum_{j \in S \setminus i} \left[g(i,j)(e_i^* + e_j^*) \right], \qquad (12)$$

where *S* defines a given protein structure. By omitting the constant factors $k_B T$ in equation (6), the resulting E_i^* are given in arbitrary unit entities and are direct proportional to energies listed in [*J*] or [kcal/mol]. [20]

The protein energy profile of a structure *S* corresponds to the n-tupel of all E_i^* usually in an interval of [-50, +10] and is, therefore a characteristic vector of energy values for each protein structure.

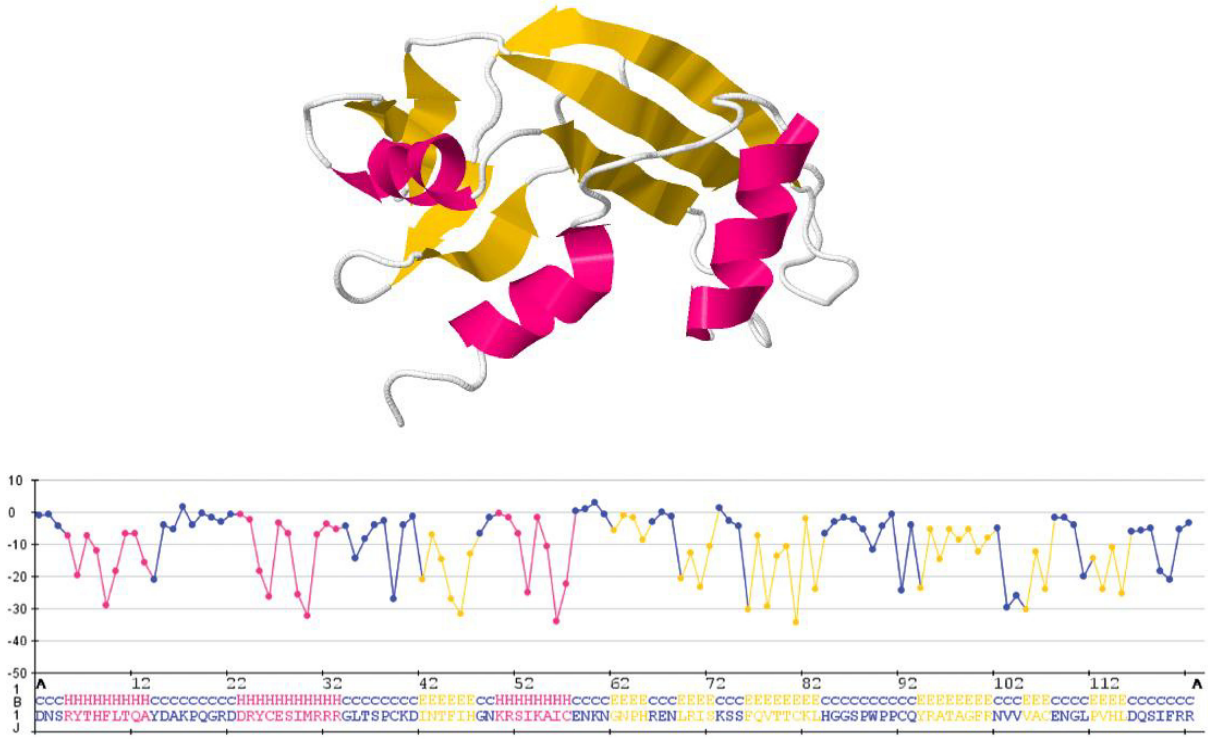


Figure 2.6: Structure of 1B1J from JMol [27] and corresponding energy profile [28]

Here, the crystal structure of the human angiogenin variant H13A (PDB ID 1b1j) and its corresponding energy profile is depicted. Coil-regions (colored blue/white) in the structure are in general, occupied by high energetic states, while the structural more stable conformations helices (pink) and strands (yellow) are featured by lower energetic states.

The eProS [4] database and toolbox provides a various number of features for the work with protein energy profiles, such as the calculation of energy profiles from globular proteins, global or local pairwise energy profile alignments and the eGOR algorithm, which aims to predict discretized energy profiles based on protein sequences.

2.2.3 Discretization of Energy Profiles

A vast amount of recent and actual work in bioinformatics is based on identified sequence motifs from multiple alignments. Algorithms like MEME [29] or The GibbsSampler [30] are focused on finding conserved motifs with biological significance from multiple sequence alignments by means of statistics. The eMOTIF database [31] contains more than 170,000 highly specific and sensitive protein sequence motifs representing conserved biochemical properties and biological functions. These protein motifs are derived from 7,697 sequence alignments in the BLOCKS [32] database and all 8,244 protein sequence alignments in the PRINTS [33] database. Other approaches on sequence alignments include the usage of position weight matrices (PWM) and their proper visualization, the Sequence Logo. [34, 35, 36]

The discovery of motifs and the generation of profiles using a set of energy profiles require the transformation of energy profiles from continuous vectors to discrete sequences. Therefore, energy profiles need to be discretizised with an approach from descriptive statistics, called n-percentilization. In general, a percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value below which 20 percent of the observations may be found.

The transformation of a continuous scaled vector of energy values into a sequence of discrete variables requires a suitable basis to calculate the n-percentiles. This basis is a set of 2,700 randomly sampled non-redundant globular protein structures. [37] Based on this structure set, where every structure has a resolution below 3 Å the corresponding energy profiles were computed and concatenated, which leads to 813,982 energy values, serving as a representative and independent set. In terms of avoiding misleading correlations, it is absolutely necessary to obtain a separate set of energy profiles to calculate the n-percentiles.

The distribution of energy values in this sampled set can be characterized with a Gaussian kernel density plot:

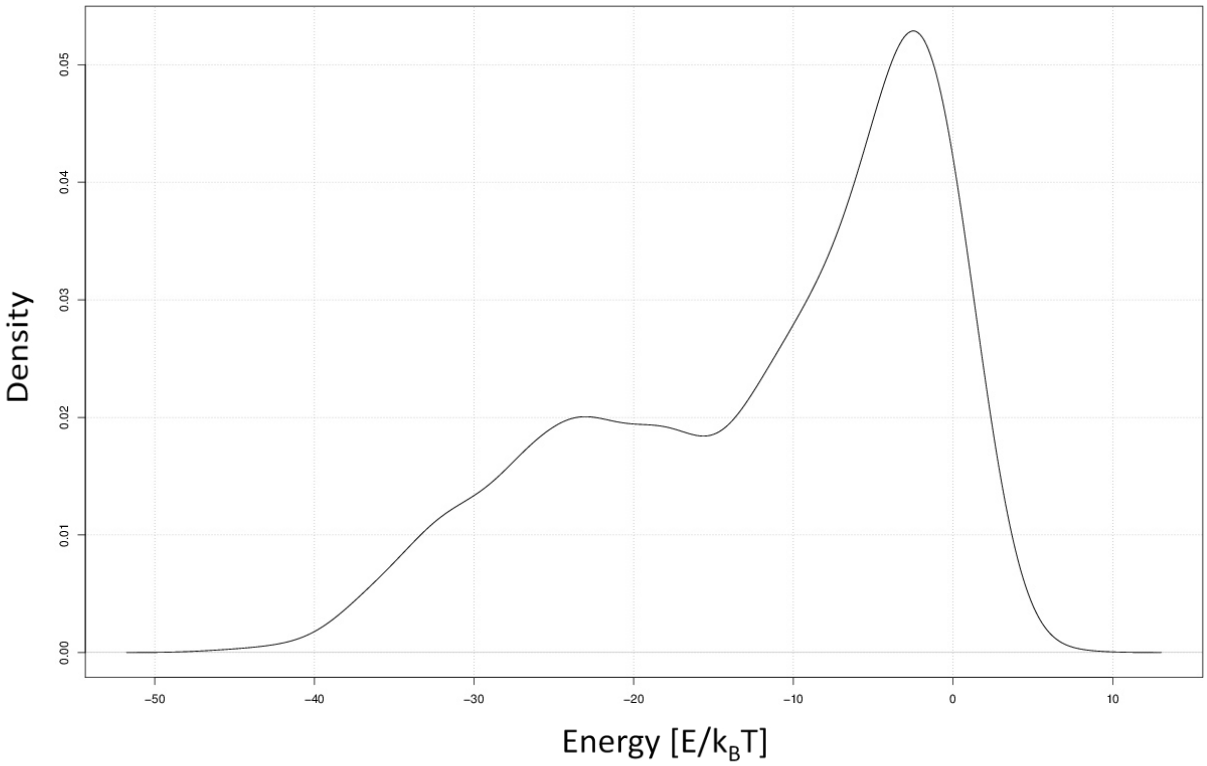


Figure 2.7: Density plot of the 813,982 energy values set

Here the distribution of the energy values from the 2,700 randomly sampled PDB-structures to calculate the n-percentiles is depicted as a Gaussian kernel density plot.

Like it has been mentioned in the chapter before, energy values are distributed among an interval of [-50, +10], but are not distributed equally in this interval. The distribution has its highest density between energy values of -10 and +2. The density of values bigger than +2 or lower than -35 is rapidly decreasing due to the fact, that such values are representing very unstable conformations or unrealistic stable states of a residue within a protein structure. The area between values of -30 and -15 represents the common case of residues, which are embedded well in a structure.

Percentiles are in general only a special case of quantiles, where quantiles are points taken at regular intervals from the cumulative distribution function of a random variable. The aim of quantile representation is to divide ordered data into *n* essentially equally sized data subsets; the quantiles are again the data values marking the boundaries between consecutive subsets. More generally, it is possible to consider the quantile function for any distribution, which is defined for real variables between zero and one and is mathematically the inverse of the cumulative distribution function.

For a population of discrete values, or for a continuous population the *k*th *n*-quantile of a random variable *X* is the data value where the cumulative distribution function crosses $\frac{k}{n}$. That is, *x* is a *k*th *n*-quantile for *X* if

$$Probability[X < x] \leq \frac{k}{n} \quad \wedge \quad Probability[X \leq x] \geq \frac{k}{n}. \tag{13}$$

For a finite population of *N* values indexed *1,…,N* from lowest to highest the *k*th *n*-quantile of this population can be computed with:

$$I_p = N\frac{k}{n}, \tag{14}$$

where I_p is the index in N of the *k*th *n*-quantile and $p = \frac{k}{n}$ is a real number within the interval $0 < p < 1$. If I_p is not an integer, it will be estimated by rounding up to the next integer to get the appropriate index. The corresponding data value to I_p is the *k*th *n*-quantile and can be referred to as the *k*th boundary of an *n*-quantile. The underlying population of values for N is the previously described 813,982 energy values containing set of randomly sampled concatenated energy profiles.

With this methodology it is possible to transform continuous energy profiles into sequences of discrete symbols. In the following deliberations these sequences are called *n*-quantile sequences, where *n* is arbitrary, but fixed.

Table 2.1: n-discretizations and their corresponding quantiles

Table 2.1 lists different *n*-discretizations and their corresponding quantiles $q_{n,m}$, where *m* is *1 < m < n-1*.

n = 4	n = 6	n = 8	n = 10	n = 12
$q_{4,1}$ = -20.21	$q_{6,1}$ =-24.74	$q_{8,1}$ = -27.16	$q_{10,1}$ = -28.73	$q_{12,1}$ =-29.82
$q_{4,2}$ = -8.43	$q_{6,2}$ = -15.66	$q_{8,2}$ = -20.21	$q_{10,2}$ = -22.88	$q_{12,2}$ = -24.74
$q_{4,3}$ = -2.94	$q_{6,3}$ = -8.43	$q_{8,3}$ = -13.52	$q_{10,3}$ = -17.47	$q_{12,3}$ = -20.21
	$q_{6,4}$ = -4.44	$q_{8,4}$ = -8.43	$q_{10,4}$ = -12.33	$q_{12,4}$ = -15.66
	$q_{6,5}$ = -1.53	$q_{8,5}$ = -5.27	$q_{10,5}$ = -8.43	$q_{12,5}$ = -11.56
		$q_{8,6}$ = -2.94	$q_{10,6}$ = -5.81	$q_{12,6}$ = -8.43
		$q_{8,7}$ = -0.8	$q_{10,7}$ = -3.81	$q_{12,7}$ = -6.19
			$q_{10,8}$ = -2.09	$q_{12,8}$ = -4.44
			$q_{10,9}$ = -0.32	$q_{12,9}$ = -2.94
				$q_{12,10}$ = -1.53
				$q_{12,11}$ = 0.0

Taking these quantiles, the energetic boundaries to an *n*-discretization and their related symbol are assigned accordingly:

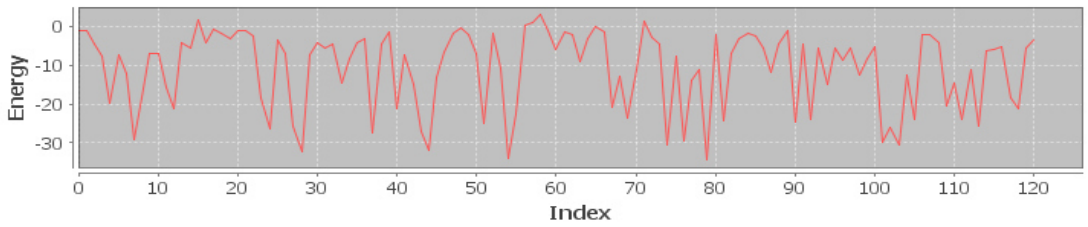
Table 2.2: Quantiles and their related symbols

Table 2.2 lists the n-quantile boundaries and the corresponding n-quantile symbols.

Symbol	n = 4 Boundaries	n = 6 Boundaries	n = 8 Boundaries	n = 10 Boundaries	n = 12 Boundaries
1	$e < q_{4,1}$	$e < q_{6,1}$	$e < q_{8,1}$	$e < q_{10,1}$	$e < q_{12,1}$
2	$q_{4,2} > e \geq q_{4,1}$	$q_{6,2} > e \geq q_{6,1}$	$q_{8,2} > e \geq q_{8,1}$	$q_{10,2} > e \geq q_{10,1}$	$q_{12,2} > e \geq q_{12,1}$
3	$q_{4,3} > e \geq q_{4,2}$	$q_{6,3} > e \geq q_{6,2}$	$q_{8,3} > e \geq q_{8,2}$	$q_{10,3} > e \geq q_{10,2}$	$q_{12,3} > e \geq q_{12,2}$
4	$e \geq q_{4,3}$	$q_{6,4} > e \geq q_{6,3}$	$q_{8,4} > e \geq q_{8,3}$	$q_{10,4} > e \geq q_{10,3}$	$q_{12,4} > e \geq q_{12,3}$
5		$q_{6,5} > e \geq q_{6,4}$	$q_{8,5} > e \geq q_{8,4}$	$q_{10,5} > e \geq q_{10,4}$	$q_{12,5} > e \geq q_{12,4}$
6		$e \geq q_{6,5}$	$q_{8,6} > e \geq q_{8,5}$	$q_{10,6} > e \geq q_{10,5}$	$q_{12,6} > e \geq q_{12,5}$
7			$q_{8,7} > e \geq q_{8,6}$	$q_{10,7} > e \geq q_{10,6}$	$q_{12,7} > e \geq q_{12,6}$
8			$e \geq q_{8,7}$	$q_{10,8} > e \geq q_{10,7}$	$q_{12,8} > e \geq q_{12,7}$
9				$q_{10,9} > e \geq q_{10,8}$	$q_{12,9} > e \geq q_{12,8}$
0				$e \geq q_{10,9}$	$q_{12,10} > e \geq q_{12,9}$
E					$q_{12,11} > e \geq q_{12,10}$
T					$e \geq q_{12,11}$

According to Table 2 every energy value *e* in an energy profile *E* can be translated into a discrete *n*-quantile symbol, depending on the choice of *n*.

Continuous scaled energy profile *E*:



813,982 energy values from 2,700 randomly sampled energy profiles for quantile calculation

Discrete n-quantile sequence of *E* with n = 4:

443323212332133434434442133113333233313413211234443142114444344234412124431312214133443234131323232331112144312

This overview depicts the transformation of continuous scaled energy profiles into discrete *n*-quantile sequences conclusively.

2.3 The Pfam Database

This chapter is dedicated to the Pfam database as a source of biological data used in this work. The approach to create a distinction of protein structures into functional families will be explained.

The Pfam database is a widely used and well-established database of protein families, containing 14,831 manually curated entries in the current release, version 27.0. [3] Each family in the database is defined by a multiple sequence alignments and a profile hidden Markov model (HMM). Profile HMMs are probabilistic models used for the statistical inference of homology [38, 39] built from an aligned set of curator-defined family-representative sequences. A high-quality seed alignment is essential, as it provides the basis for the position-specific amino acid frequencies, gap and length parameters in the profile HMM. [3] The central idea behind this distinction is that sequence identity over a certain level indicates homology and therefore evolutionary relationships between protein sequences. Proteins are generally composed of one or more functional regions, commonly termed domains. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function. [40]

The families provided by the Pfam database can be distinguished into two subgroups: Highly qualitative and manually generated Pfam-A families and automatically assigned Pfam-B families. The assignment of a protein sequence to a Pfam-A family is following four general steps [1]:

I. Manually generation of a multiple sequence alignment, the so called seed-alignment from a non-redundant representative set of full-length domain sequences surely belonging to the family [1]
II. Construction of a profile HMM using the HMMER3 software-suite [41, 42]
III. Search of the profile HMM against a large UniProtKB database derived sequence collection to find all instances of the family
IV. Calculation of family-specific sequence and domain thresholds, called gathering thresholds – sequence regions that score above the required threshold that is set for each family to eliminate false positives are aligned to the profile HMM to produce the full alignment

The initial members of a seed are collected from one of several sources: Swissprot [43], Prosite [44], structural alignments [45], ProDom [46], BLAST results [47], repeats found by Dotter [48] or published alignments. With automated alignment methods like ClustalW [49], ClustalV [50] or HMM training [51] the seed alignment is generated. This alignment needs to be as qualitative as possible.

Originating from a seed alignment an HMM is built by using the HMMER3 [41] software-suite. To avoid overfitting and to design the HMM more universal, amino acid frequency priors were derived according to an ad-hoc pseudocount method using the BLOSUM62 [52] substitution matrix. Each constructed HMM is then compared with all sequences in Swissprot to gather all instances of a family. [1] By extracting all matching sequence fragments and aligning them to the HMM, the final full alignment is created.

In some cases, a profile HMM is not able to detect all homologues of a diverse superfamily, so multiple entries may be built to represent different sequence families in the superfamily. Such related Pfam-A entries are grouped into clans. [53] Although these Pfam-A entries cover a large fraction of the sequences in the underlying sequence database, in order to give a more comprehensive coverage of known proteins Pfam generates a supplement using the ADDA [54] database. These automatically generated entries are the Pfam-B families. Although of lower quality, Pfam-B families can be useful for identifying functionally conserved regions when no Pfam-A entries are found.

Conclusively a Pfam entry represents a family of functionally related protein-segments. Those segments or conserved regions are referred to as domains. For instance, if one is searching for the term ‘Lysozyme’ in Pfam the database will lead the user to the clan or superfamily ‘Lysozyme-like superfamily’ with the Clan ID CL0037, which contains 33680 sequences from 5,612 different species. [36] The clan in terms contains the following 12 member families:

- Glucosaminidase
- Glyco_hydro_108
- Glyco_hydro_19
- Glyco_hydro_46
- Lys
- Lysozyme_like
- Phage_lysozyme
- REGB_T4
- SLT
- SLT_2
- TraH_2
- Transglycosylase

By picking out one of these families, Lys for instance, the exact family name '**C-type lysozyme/alpha-lactalbumin family**' and its referring Pfam ID PF00062 are returned. The family PF00062 in terms consists of 9 domain organizations found in 1,035 sequences:

Table 2.3: The Pfam family PF00062

Table 2.3 lists the different domain organizations for PF00062 and the number of sequences representing a particular organization.

Domain organisation	Number of representing sequences
	897
	27
	13
	4
	1
	1
	1
	1
	1
	1

Based on this the user may create a sequence alignment or view some structure for further work with the specified family. The Lys-family PF00062 will serve as an example in some of the following chapters to demonstrate some techniques used in this work exemplarily.

2.4 Methods of Sequence Discovery

This last methodology chapter is dedicated to fundamentally methods of sequence discovery as Position Weight Matrices (PWMs) and their derived visualization as sequence logos. These techniques are essentially inspired by common protein substitution matrices and information theory and hence, provide a valuable way to approach discretizised energy profiles.

2.4.1 Position Weight Matrices

The position weight matrix (PWM), also known as position-specific weight or scoring matrix is a commonly used representation of motifs and regular patterns in biological sequences. PWMs are often referred to as profiles of a set of sequences and therefore as a motif descriptor. It attempts to capture the intrinsic variability characteristic of sequence patterns. Derived from multiple sequence alignments the PWM represents the positional dependency of symbol-distributions among a set of sequences.

Introduced by the American geneticist Gary Stormo in 1982 [6] the PWM is an alternative to consensus sequences. Consensus sequences had previously been used to represent patterns in biological sequences, but had difficulties in the prediction of new occurrences of these patterns. [51] The first application of PWM's was the discovery of RNA sites that function as translation initiation sites. [6] Derived from the weighting vectors used in the perceptron algorithm the PWM was suggested by Polish mathematician Andrzej Ehrenfeucht in order to create a matrix of weights which could distinguish true binding sites from other non-functional sites with similar sequences. The training process of the perceptron on both sets of sites resulted in a matrix and a threshold to distinguish between the two sets. Stormo used this matrix to scan new sequences not included in the training set to assign the desired sites correctly and proved, that this method was both more sensitive and precise than the use of the best consensus sequence. [6] Many current algorithms and applications are using PWM's over consensus sequences to represent patterns in biological sequences. [52, 53]

Algorithms like MEME or The GibbsSampler are focused on finding conserved motifs with biological significance from multiple sequence alignments. Those pattern searching algorithms assume that the user needs to know the concrete form of a pattern or motif. In contradiction position weight matrices contain all occurring patterns among a set of aligned sequences; weighted by their respective frequency.

The values assigned to a PWM are both position and symbol specific log-likelihoods depending on an underlying probabilistic background model. The simplest background model assumes that each symbol appears equally frequented in the given dataset, which will lead to background frequencies 0.25 for nucleotides and 0.05 for amino acids. Although the applicability of such a simple model is possible, the probabilistic model to calculate the background frequencies used in this work is different. Originating from a multiple sequence alignment of N_{seq} a PWM consists of L_{aln} columns with each r rows, where L_{aln} is the position wise length of the alignment and r the number of symbols occurring in the given alphabet A. Common alphabets are those for DNA with $r = 4$ and proteins with $r = 20$. Obviously the number of distinct symbols and therefore the dimensions of the PWM are depending on the data basis, which is why the shape of the PWM's used in this work depends on the n of an n-discretization of a set of energy profiles.

The starting point for calculating PWM values is the relative frequency of a symbol a occurring in the uth alignment column:

$$f_{u,a} = \frac{n_{u,a}}{N_{seq}} \qquad (15)\ ,$$

where $n_{u,a}$ is the absolute frequency of symbol a occurring in alignment column u. If the PWM is calculated from multiple protein alignment sequences the value associated with row b and column u will be:

$$m_{u,b} = \sum_{\forall a \in A} f_{u,a}\, s_{a,b}\ . \qquad (16)$$

Here, $s_{a,b}$ is the substitution value of residue a to residue b derived from a given protein substitution matrix like BLOSUM-62 [54] for example. Due to the lack of those substitution values applicable to n-discretizised energy profiles the following log odds form for a PWM element $m_{u,a}$ will be used:

$$m_{u,a} = \log \frac{q_{u,a}}{p_a}\ . \qquad (17)$$

If there are sufficient sequence data available $q_{u,a}$ can be identified with $f_{u,a}$ as given by equation (15). [55] Here, p_a is the background frequency of *a* derived from an adequate underlying probabilistic model.

The log-odd scheme for calculating the PWM elements $m_{u,a}$ given by (17) does have one main drawback: if there is no occurrence of *a* symbol *a* in column *u* the PWM element $m_{u,a}$ will be assigned with an undefined value, due to the fact that *log(0)* is mathematically not defined. [55] Such an absence of a particular symbol type in a particular column of the alignment used to derive the PWM is more likely to indicate a lack of data rather than the true symbol preferences. Hence, this problem occurs due to a lack of underlying alignment data and has to be solved.

A simple way to overcome this lack of data is to assume at least one occurrence of each symbol type at each alignment position. Therefore the calculation of $q_{u,a}$ used in equation (17) has to be adopted in a way, that all symbols and positions are treated equally:

$$q_{u,a} = \frac{n_{u,a} + 1}{N_{seq} + |A|}, \qquad (18)$$

where |A| describes the number of distinct symbols contained in a given alphabet A and is also referred to as cardinality. The sum of the $q_{u,a}$ for all possible symbols $a \in A$ must always equal 1 as each sequence must be represented at column u. The inclusion of the additional observation means that $q_{u,a}$ will never be 0 nor ever reach 1. Such additional data is often referred to as pseudocounts. [51] Although it is possible to calculate useful PWM values with this scheme there are more sophisticated ways of adding pseudocounts that take advantage of the knowledge of the properties of sequences. Again, the PWM value calculation done by equation (18) assumes that the symbol distribution of a given alphabet is uniform. This assumption is deniable for protein alphabets as well as alphabets originating from *n*-discretizised energy profiles, which is why the influence of the pseudocounts to calculate a PWM value can be adopted:

$$q_{u,a} = \frac{n_{u,a} + \beta p_a}{N_{seq} + \beta}. \qquad (19)$$

Here the parameter β is a scaling parameter that determines the total number of pseudocounts in an alignment column. The advantage of introducing such a scaling parameter is that it is easily possible to adjust the weighting of real data and pseudocounts.

When there is a lot of underlying alignment data and N_{seq} is large there is little if any need for pseudocounts and β should be larger than N_{seq}, whereas when there is less data, β should be larger relative to N_{seq}. A simple approximation for β has been found [65]:

$$\beta = \sqrt{N_{seq}} \qquad (20)$$

At large values of N_{seq} equation (18) and (19) approaches $q_{u,a}$ and $f_{u,a}$ as desired. In the absence of any data the pseudocounts would completely determine the PWM values and in terms of Bayesian analysis they are representing the prior distribution, which expresses the prior knowledge of the system before any data is introduced to the calculation. The values assigned to a PWM are often referred to as log-likelihoods.

Once a PWM has been derived from a set of functionally related sites, it can be used to scan a query sequence for the presence of potential sites or characteristic patterns usually by running a window of the matrix along the sequence and sum the position weight values from corresponding to each symbol in each position on the window sequence. [56]

Formally, the score S_q of a query sequence q of length $|q|$ consisting of distinct symbols $a \in A$ can be calculated as the sum of the particular n-qPWM values $m_{u,a}$:

$$S_q = \sum_{u=0}^{|q|} m_{u,a}. \qquad (21)$$

This score is a quantitative measurement whether the underlying set of sequences and the query sequence have profile characteristics in common or not. [56]

2.4.2 Sequence Logos

Sequence logos are an often used graphical representation of the sequence conservation, created from a set of aligned DNA- or protein sequences to extract and visualize both consensus sequences and sequence patterns. They were invented by Tom Schneider and Mike Stephens in 1990 [57] to display patterns in sequence conservation, and to assist in discovering and analyzing protein or DNA-binding sites. The importance of a particular position in a binding site is clearly and consistently given by the information required to describe the pattern there. [58, 59]

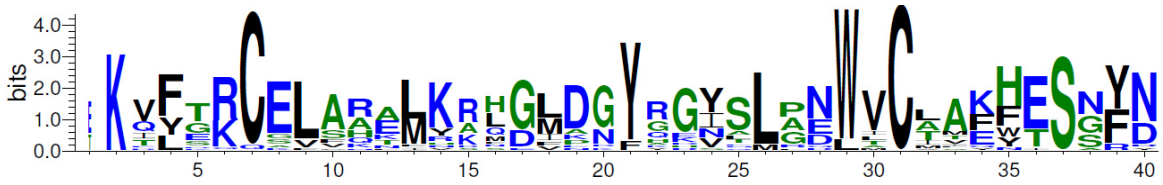


Figure 2.8: Sequence Logo of Lysozyme family PF00062

Here the first 40 positions of 25 sequences of the Lysozyme family PF00062 are depicted as sequence logo generated by the Berkeley Sequence WebLogo [5] Generator.

Every particular position in the logo represents the information content at this position measured in bits. This information content or total information *R* at a particular position *i* is defined as the difference between the maximum uncertainty (equation 23) and the uncertainty of the observed symbol distribution (equation 24) at this position [56]:

$$R_i = H_{max} - [H_{obs} + e(n)] \qquad (22)$$

$$H_{max} = \log_2 |A| \qquad (23)$$

$$H_{obs} = -\sum_{\forall\, a \in A} f_{a,i} \log_2 (f_{a,i}) \qquad (24)$$

Again *A* is the alphabet of the underlying data and *|A|* its cardinality. The factor $f_{a,i}$ describes the relative frequency of a symbol *a* at position *i* and *e(n)* (see equation 25) is a correction factor required when only a few sequences are available. [60].

This sampling error correction is the expectation of a sampled uncertainty depending on the number of sequences used to create the logo and can be approximated with [61, 62]:

$$e(N_{Seq}) = \frac{|A| - 1}{2 \ln(2)\, N_{Seq}} \qquad (25)$$

Finally the height of a symbol at a particular position is calculated as:

$$height_{a,i} = f_{a,i}\, R_i\ . \qquad (26)$$

The height of the entire stack of symbols is proportional to the information content at that position. Consequently, the maximum uncertainty per site is $log_2\ 4 = 2$ bits for DNA/RNA and $log_2\ 20 \approx 4.32$ bits for protein sequences. The uncertainty of the observed symbol distribution at a particular position calculated with equation (24) is often referred to as the uncertainty measure. [57]

Applying the sequence logo approach to *n*-discretizised energy profiles the maximum uncertainty will depend on the choice on *n*. Neglecting the inter-site correlations and assuming a uniform background symbol distribution, then the total entropy of the logo, the sum of the sequence conservation at each position, measures the information content of the logo. [5] For binding sites, this total entropy has, in many cases, been shown to be approximately equal to the amount of information needed to locate the binding site within the relevant stretch of DNA. [63]

3 Pfam Dataset Construction

In this chapter the construction of the datasets used in this work will be described. To ensure that the following work with energy profiles and pfam family alignments return significant results the data basis needs to be representative and free of redundancies.

3.1 Preparation of the Dataset

The first fact that limits the number of usable data for the work with the pfam families is that they are generated from multiple sequence alignments while energy profiles are structure based. Not every protein that has been sequences, possesses a corresponding 3D structure available in the PDB. Hence, only family entries with a determined 3D structure can be used. The pfam database provides a listing of all family entries with a determined 3D structure. Furthermore it contains a mapping of the the start and end positions of 197,858 domains onto the set of available 3D structures in the PDB.

A first criterion for the generation of a suitable family dataset was the exclusion of membrane proteins due to the lack of an energy function for those structures. The energy profile approach as outlined in section 2.2.2 is only applicable to globular proteins. The next criterion was the availability of a determined 3D structure in a pfam family. Only families with sequences with at least one available structure from the PDB were included. Finally there are 34,483 sequences from 5,857 families contained in the dataset. The next and most important step was the mapping of the pfam entries to their related energy profile fragments.

One main problem has occurred during this mapping step: the availability of a determined 3D structure does not indicate that an energy profile calculation on this structure is feasible. This is due to missing regions or smaller gaps within a protein structure caused by crystallization errors in the process of structure determination. Those erroneous regions would have caused gaps or even frameshifts in the energy profile mapping leading to possible non-remediable errors later.

To decide whether the mapping of an energy profile to a given Pfam family entry does make sense or not the sequences contained in an energy profile were compared to the sequences from Pfam using the Needleman-Wunsch alignment algorithm [64]. The next steps, including the calculation of Needleman-Wunsch scores, their normalization as well as their evaluation to solve the problem outlined above, were done by members of the Bioinformatics Group Mittweida [65].

At first it is necessary to verify whether a Needleman-Wunsch (NW) alignment of the energy profile sequence EP_{seq} and the Pfam entry sequence Pf_{seq} is random or not. Therefore a raw NW-Score Y_{raw} between the two obtained sequences was used as decision criterion:

$$Y_r = NWScore\left(EP_{seq}, Pf_{seq}\right) \qquad (27)$$

To estimate a relevant threshold for Y_{raw} a mean expected score Y_{mean} is calculated by:

$$PermSum = \sum_{i=1}^{100} Y_{raw}\left(Perm\left(EP_{seq}\right), Pf_{seq}\right) \qquad (28)$$

$$Y_{mean} = \frac{PermSum}{100}. \qquad (29)$$

Here P*ermSum* is the sum of scores of 100 random permutations $Perm(EP_{seq})$ of the energy profile sequence and the Pfam family entry sequence Pf_{seq}. Another score which has to be included is Y_{opt}:

$$Y_{Opt} = NWScore\left(Pf_{seq}, Pf_{seq}\right) \qquad (30)$$

Y_{opt} describes the optimal score of a Pfam family entry sequence aligned with itself. Originating from Y_{raw}, Y_{mean}, and Y_{opt} the real weighted score Y_{real} is calculated by means of:

$$Y_{real} = -\log\left(\frac{Y_{raw} - Y_{mean}}{Y_{opt} - Y_{mean}}\right) \qquad (31)$$

This final real weighted score Y_{real} now may serve as a decision criterion whether to refuse an alignment between EP_{seq} and Pf_{seq} or not. Additionally a comparative value to the final score has to be found. Therefore the correlations between Y_{real} and other parameters governed by the alignment were discovered. The sequence identity as one of those alignment parameters was revealed as the best parameter correlating with Y_{real}. With a determination coefficient $R^2 = 0.9821$ the correlation was proved as significant.

Considering the regression function supplied by the correlation between sequence identity and Y_{real} a predictive score Y_{pred} can be approximated:

$$Y_{pred} = -1.006 \ln(SeqID) + 4.7189 \qquad (32)$$

By means of standardization through z-transformation it is possible to determine the final decision property:

$$t = \frac{Y_{real} - Y_{pred}}{StdDev} \qquad (33)$$

Here StdDev is with a value of 0.03858 the standard deviation of the correlation mentioned above. An alignment between EP_{seq} and Pf_{seq} will be accepted if the predictive score Y_{pred} lies within the observed dataspace with a probability bigger than 95%. The threshold value t is delivered by this probability and the standard normal distribution table of the normal distribution integral. By claculating the area under the graph of the standard normal distribution t can be determined as 1.65.

This means, if an alignment between an energy profile sequence and a Pfam family entry sequence returns a t-value bigger than 1.65 it will be accepted as a non-random significant alignment. Applying these steps on the whole dataset 30,543 sequences in 5,082 Pfam families remained in the set. Finally only those families that contain more than 10 entries in the form of aligned sequences shall be obtained, which results in a final number of 743 families in the whole Pfam family dataset. Note, that these families are containing about 2/3 of all alignments.

3.2 Formats and Alignment Energy Mapping

With the information delivered by the Pfam database, where a particular family domain is located in a protein structure it is possible to map the corresponding energy values onto a set of family sequences. It is remarkable, that the family domain information from Pfam conveys only knowledge about the position and the sequences of structure fragments. Therefore only fragments of whole energy profiles are mapped onto structures containing family domains, which are in turn also fragmentized.

```
>LYSC1_CANFA/1-127:1EL1-A/1-127
-KIFSKCELARKLKSMGMDGFHGYSLANWVCMAEYESNFN
>LYSC2_ONCMY/16-142:1BB6-A/1-127
-KVYDRCELARALKASGMDGYAGNSLPNWVCLSKWESSYN
>LYSC1_HORSE/1-127:2EQL-A/1-127
-KVFSKCELAHKLKAQEMDGFGGYSLANWVCMAEYESNFN
>LYSC_MELGA/19-145:135L-A/1-127
-KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFN
>LYS_BOMMO/19-137:1GD6-A/1-119
-KTFTRCGLVHELRKHGFE---ENLMRNWVCLVEHESSRD
>LALBA_PAPCY/1-120:1ALC-A/1-120
-KQFTKCELSQNLY--DIDGYGRIALPELICTMFHTSGYD
>Q7YT17_MUSDO/1-121:3CB7-A/5-125
-KTFTRCSLAREMYKLGVP---KNQLARWTCIAEHESSYN
>LYSC_COTJA/19-145:2IHL-A/1-127
-KVYGRCELAAAMKRHGLDKYQGYSLGNWVCAAKFESNFN
>LYSC_COLVI/1-127:1DKJ-A/1-127
-KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFN
>LALBA_BOVIN/20-139:2G4N-A/1-120
E-QLTKCEVFRELK--DLKGYGGVSLPEWVCTTFHTSGYD
>LYS_ANTMY/1-120:1IIZ-A/1-120
-KRFTRCGLVNELRKQGFD---ENLMRDWVCLVENESARY
>LYSC_NUMME/1-127:1HHL-A/1-127
-KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFN
```

Figure 3.1: Pfam alignment energy mapping data format

Originating from Pfam sequence alignments the whole set of family entries has been mapped onto their particular corresponding energy values. The colorings of the residues are representing the four percentiles of a 4-discretization of the energy values, where blue is the first (energetic lowest), green the second, orange the third and red the fourth (energetic highest) percentile.

In order to be consequent figure 3.1 depicts a segment of the alignment of Lysozyme family PF00062. Comparing this energy mapping to the sequence logo in figure 2.10 from section 2.4.2 it is easy to see the family-characterizing residues conserved in these sequences.

Furthermore some energetic conservations can also be observed already at this state. For instance, the lysine K at the second position in figure 3.1 seems to be highly conserved on the sequence level as well as in its energetic behavior. Another important example in this energy mapped alignment is the blue block in the right part of figure 3.1 starting with tryptophane W. While the residue distribution in this block differs in every sequence the energy belongs in nearly all cases to the first percentile. Such conspicuous issues will be extensively discussed later.

4 Approaches

This chapter melds all the methodology and knowledge from the sections before to describe some approaches for finding energetic regularities within the energy-mapped Pfam alignments: a primitive naive motif finder, a position weight matrix derived from *n*-discretizised energies, the calculation of energy logos and the development of a conservation measurement to compare sequence and energy logos. Described are only the individual steps of the different approaches. The results they delivered are presented in the next chapter.

4.1 Primitive Motif Finder

The first approach to discover energetic patterns within a set of energy mapped Pfam alignments was a primitive naive motif finder. This finder was based on the assumption, that some positions in a Pfam alignment are energetic conserved as they are on the sequence level. If a particular position in an alignment is realized by a sufficient amount of the same percentile over a set of sequences this position will be counted as a conserved one. The procedure to find these possible "conserved" motifs was as following:

I. Initialize *n* and $t \in [0, \dots, 100]$

II. Obtain all *n*-quantile sequences from a given family alignment

III. Create a symbol occurrence map of the form $|A|\ x\ L_{algn}$:

$$\begin{pmatrix} a_1 \rightarrow \left[o_1, \dots, o_{L_{algn}}\right] \\ a_2 \rightarrow \left[o_1, \dots, o_{L_{algn}}\right] \\ \dots \rightarrow \quad \dots \\ a_T \rightarrow \left[o_1, \dots, o_{L_{algn}}\right] \end{pmatrix}, \quad a_1, \dots, a_T \in A$$

IV. Parse over all n-quantile sequences and assign position- and symbol-wise occurrences $o_1, \dots, o_{L_{algn}}$

V. Iterate over L_{algn} and concatenate every symbol that has an occurrence bigger than *t* to the consensus sequence, otherwise concatenate a wildcard symbol

The result of this simple occurrence counting and concatenation process is a sequence of n-quantile symbols that are believed to be conserved. In order to have a reference the same procedure was done with residue sequences. As an example both 4-quantile and residue motif were computed from the 25 entries in the Lysozyme family PF00062 with a threshold t = 80. The first 40 positions of these motifs are shown below:

4-quantile motif :	-4--3---1---1----1-------122111-1------
Residue motif:	-K----C-L----------Y----L--W-C-----S--

The threshold value t = 80 means, that every residue or 4-quantile symbol which occurs in at least 80% at a particular position is assigned to the motif. By comparing the residue motif with the sequence logo shown in figure 2.10 from section 2.4.2 the conserved positions can be easily recognized here.

4.2 Position Weight Matrices from n-discretizised Energy Profiles

As outlined in section 2.4.1 position weight matrices (PWMs) originating from *n*-Percentile alphabets were computed. To separate the matrices used in this work from their more general ancestors as described in section 2.4.1 they are now referred to as *n*-quantile position weight matrices or simply *n*-qPWMs. The procedure in this special case of a data basis differs only in the fact that the dimension of the *n*-qPWM depends on the cardinality of the chosen n-quantile alphabet.

The elements $m_{u,a}$ assigned to the *n*-qPWM are again calculated by:

$$m_{u,a} = \log \frac{q_{u,a}}{p_a} \qquad (17)$$

$$q_{u,a} = \frac{n_{u,a} + \beta p_a}{N_{seq} + \beta} \text{ , with } \beta = \sqrt{N_{seq}} \qquad (18)$$

and $n_{u,a}$ as the absolute frequency of symbol *a* occurring in alignment column *u*, p_a as the background frequency this symbol and N_{seq} as the number of sequences contained in the alignment. The relative background frequencies *p* are derived from an underlying probabilistic model, which covers all *n*-quantile sequences of all families contained in the Pfam dataset.

The following table lists the relative background frequencies p_n of all *n*-quantile symbols used as the underlying probabilistic model for the *n*-qPWM calculation:

Table 4.1: Background frequencies of n-quantile symbols at different values of n

Table 4.1 lists the background frequencies p of the n-quantile symbols for different n-discretizations.

Symbol a	p_4	p_6	p_8	p_{10}	p_{12}
1	0.288	0.200	0.154	0.126	0.107
2	0.256	0.173	0.346	0.111	0.092
3	0.238	0.171	0.126	0.103	0.089
4	0.218	0.163	0.130	0.100	0.085
5		0.146	0.125	0.105	0.084
6		0.146	0.113	0.100	0.088
7			0.108	0.093	0.084
8			0.110	0.087	0.079
9				0.086	0.074
0				0.089	0.072
E					0.071
T					0.075

By using these background frequencies derived from the whole family dataset and applying the equations (17) and (18) to on the Pfam family PF00062 the final position weight matrix can be calculated.

4.3 Energy Logo

In order to visualize the diversity of n-quantile sequences within a particular Pfam family alignment or a set of sequences originating from a folding class as provided by CATH an equivalent to the sequence logo approach – the energy logo was realized. The procedure of generating an energy logo instead of a sequence logo is equivalent to the approach described in section 2.4.2, except that the dimension of the maximum possible uncertainty H_{max} depends on the choice of n for an n-discretization of energy profiles.

$$H_{max} = \log_2|A| = \log_2(n) \qquad (34)$$

For instance the maximal possible uncertainty for a 4-quantile logo is 2 bits, while there are 3 bits for an 8-quantile logo. However, this circumstance does not affect the meaningfulness of a logo; it is an important fact to recognize. Analogous to the calculation of symbol heights in the sequence logo the height of a particular n-quantile symbol a at position i in the energy logo can be calculated by:

$$height_{a,i} = f_{a,i}\left(\log_2 n - \left[-\sum_{\forall\, a \in A} f_{a,i}\log_2(f_{a,i}) + e(N_{Seq})\right]\right) \qquad (35)$$

The energy logo below depicts the symbol distribution of all 4-quantile sequences of the first 40 positions of the Lysozyme family PF00062 alignment:

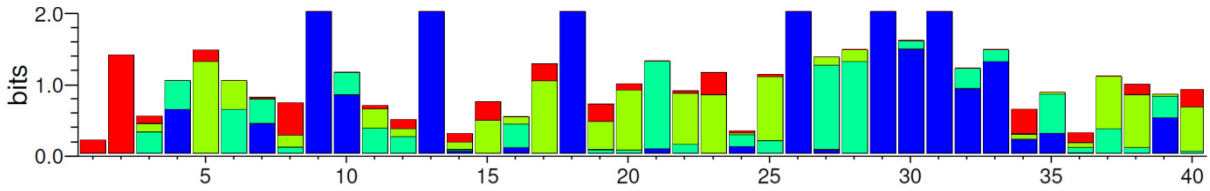


Figure 4.1: 4-quantile energy logo of PF00062

Instead of coding the symbols as letters the developed energy logo encodes the n-quantile symbols as color bars. A bar with a maximum uncertainty of 2 bits indicates that this particular position is realized by exact one symbol in all n-quantile sequences.

Considering the same sequences from the same family with an 8-quantile discretization the symbol distributions are more compensated:

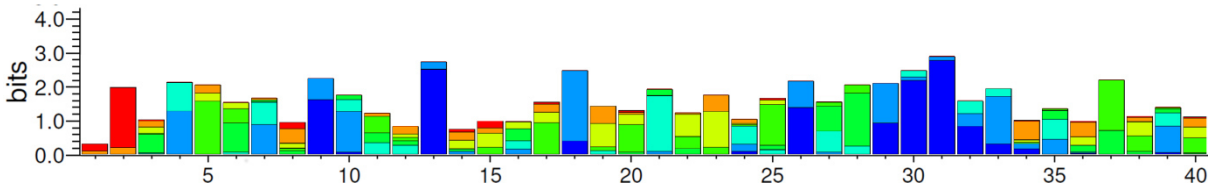


Figure 4.2: 8-quantile energy logo of PF00062

By dividing the continuous energy scale into more quantiles the generated energy logo is getting are higher resolution regarding to the distribution of the *n*-quantile symbols.

By comparing the 4-quantile and the 8-quantile energy logo of the first 40 position of the Lysozyme family PF00062 alignment it is obvious that at increasing *n*'s the maximum bar height relative to the maximum uncertainty measure H_{max} decreases. With increasing *n* the symbol distribution of an n-discretization becomes also more uniform, so that the whole shape of the 8-quantile energy logo seems to be less meaningful. However, by considering the maximum uncertainty (2 bits for 4 quantiles and 3 bits for 8 quantiles) shows that this assumption has to be denied. The bar-coloring used in the energy logo representation is derived from a continuous color-spectrum representing the energy scale:

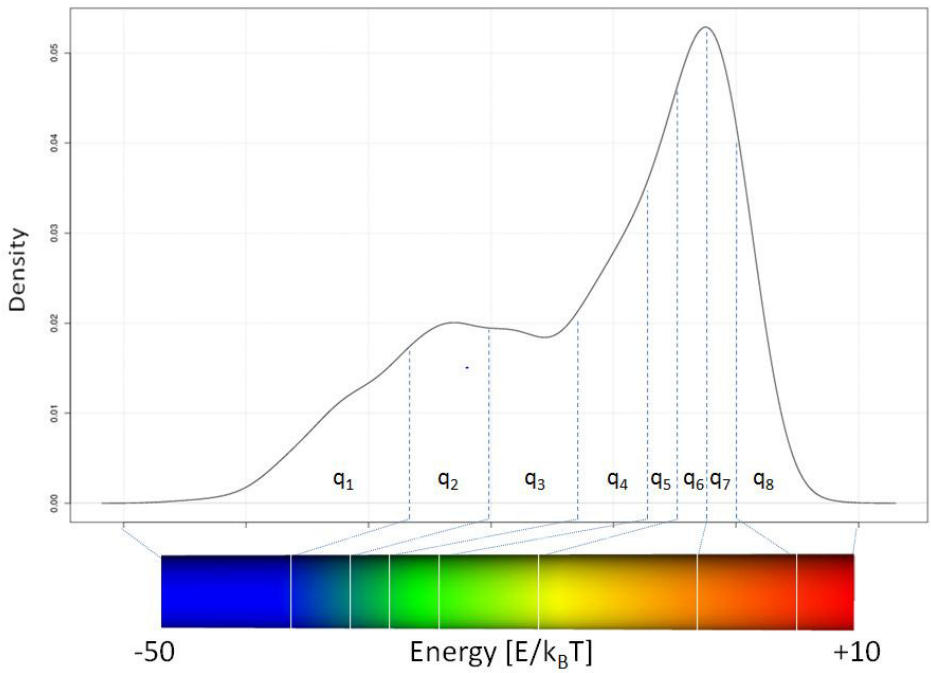


Figure 4.3: Coloring scheme of the energy scale in energy logos

A continuous color spectrum was used to colorize the n-quantile symbol bars in the energy logo referring to their specified energy interval as it is listed in table 2.1.

By comparing these energy logos to the sequence logo of the same family alignment segment as shown in section 2.4.2 some really interesting connections between sequence and energy are revealed, which will be discussed later.

4.4 Energetic & Sequence Conservation

The recent comparisons between sequence and energy logos regarding to their conservations are done by visual judgment, which is a barely unscientific way to evaluate the connections between both methods. Therefore a more sophisticated and founded method to evaluate the different logos has been developed. Originating from the height-calculation used in both logo types (see equation 35) the conservation *C* at a particular position *i* can be calculated via cumulative normalized height values:

$$C_i = \frac{\sum_{\forall a \in A} height_{a,i}}{height_{a,i}} \qquad (36)$$

Here the sum of the particular symbol heights at a given position *i* is divided with the maximum possible height, that might occur in a logo. This maximum possible height $height_{max}$ is calculated by concatenating an arbitrarily character at the end of every sequence in the alignment and determine its height. This additional character is not included in the visual logo representation. The value of $height_{max}$ depends on N_{seq} - the number of sequences contained in the alignment - and can be described with an exponential increasing function:

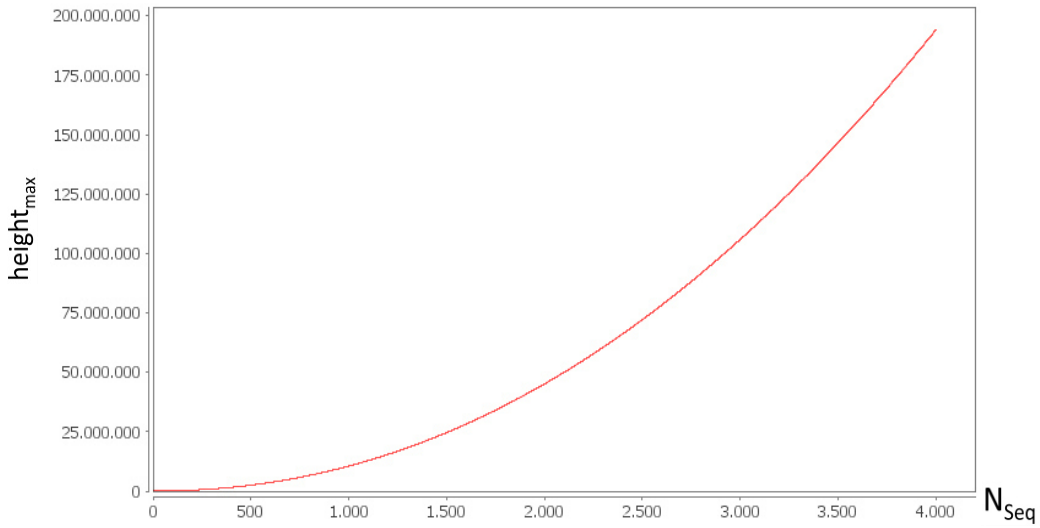


Figure 4.4: Dependency of *height*$_{max}$ and *N*$_{Seq}$

Here the values for *height*$_{max}$ of 4000 different numbers of sequences *N*$_{Seq}$ are plotted. These 4000 distinct amounts are [1, 2, … , 4000], which means that the value pairs are of the form (*N*$_{Seq}$ -> *height*$_{max}$). It is obvious, that the number of sequences and *height*$_{max}$ have a significant correlation.

Finally the conservation comparison between a sequence and an energy logo via the cumulative normalized height values is visualized below:

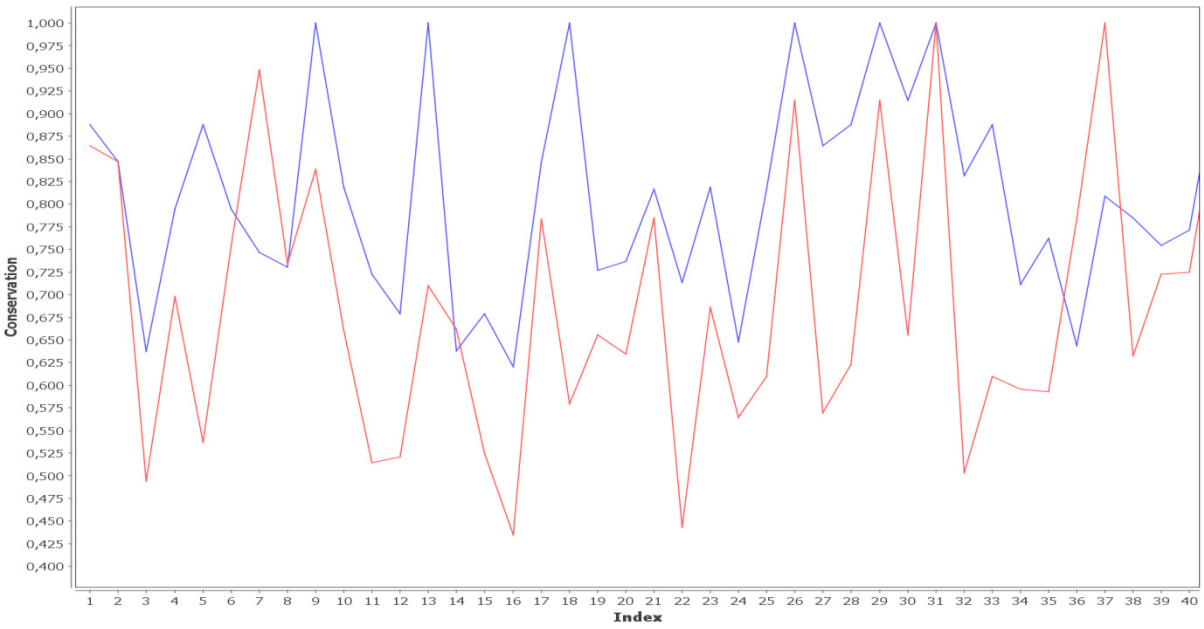


Figure 4.5: Cumulative normalized height curves

In this graph the cumulative normalized height values of both the sequence logo shown in section 2.4.2 (red curve) and the 4-quantile energy logo (blue curve) depicted in Figure 4.1 are plotted.

The conservation measurement via the cumulative normalized heights depicted in Figure 4.3 reveals, that the information content regarding to conservation differs at some particular positions significantly. Such positions, which appear as highly conserved in one logo, whereas they are containing lesser information in the other logo will be discussed in the next section.

Note, that the plotting of the cumulative normalized heights of two logos does not provide more information as available in the logos itself. The calculation of the cumulative normalized heights is just a measurement to compare both logo-types on a shared kind of scale.

5 Results and Discussion

In this last chapter the results gained through the approaches completed in the section before will be summarized and discussed. While the results of the primitive motif finder approach and the n-qPWM approach will be handled short, the comparison and interpretation of sequence and energy logos will get the focus.

5.1 Primitive Motif Finder

To evaluate the results given by the primitive motif finder approach as outlined in section 4.1 some common measurements regarding the quality of a classifier were applied. The primitive motif finder defines an n-quantile motif within a Pfam family alignment as a consensus sequence of n-quantile symbols, where every symbol in the consensus sequence occurs with a relative frequency over a given percentage threshold among all alignment sequences. Therefore, the form of a particular n-quantile motif depends on n and the given threshold and every family alignment can be assigned with an n-quantile motif. To measure the quality of those n-quantile motifs the generated motif was interpreted as a regular expression and was ran naively against the whole Pfam family dataset to extract true positive, true negative, false positive and false negative values.

A true positive hit for instance is counted, if any n-quantile family sequence contains the specified n-quantile motif and furthermore belongs to the family, who defines the motif. Obviously this test should only reveal how reliable an n-quantile motif generated by the primitive motif finder describes its family.

The Pfam family PF00069 alignment contains 187 entries and is, therefore one of the families with the biggest amount of sequences in the dataset. To ensure, that the derived motifs are generated with an adequate data basis this family is chosen to discuss some results of the primitive motif finder exemplarily.

Like it has been mentioned in section 4.1 both motif-types - n-quantile motif and sequence motif - are generated simultaneously to reference to each other. In order to describe the behavior of the primitive motifs depending on their percentage threshold the F_1 measure of 35 4-quantile and 35 analogue constructed amino acid motifs derived from the PF00069 family alignment were calculated with:

$$F_1 = \frac{2(precision * recall)}{precision + recall} \quad (37)$$

$$precision = \frac{t_P}{t_P + f_P} \quad (38)$$

$$recall = \frac{t_P}{t_P + f_N} \quad (39)$$

Here *precision* or the *positive predictive value* is the ratio between correct classified sequences measured among all sequences that are matching to a motif. The *recall* or *sensitivity* is the ratio between correct classified sequences and all sequences that are truly belonging to the Pfam family PF00069. Finally, the F-measure is a combined quality measure for classifiers, which involves precision and recall and weights them equally via the harmonic mean. Plotted below are the F_1 values for 35 different 4-quantile and residue derived motifs from Pfam family PF00069:

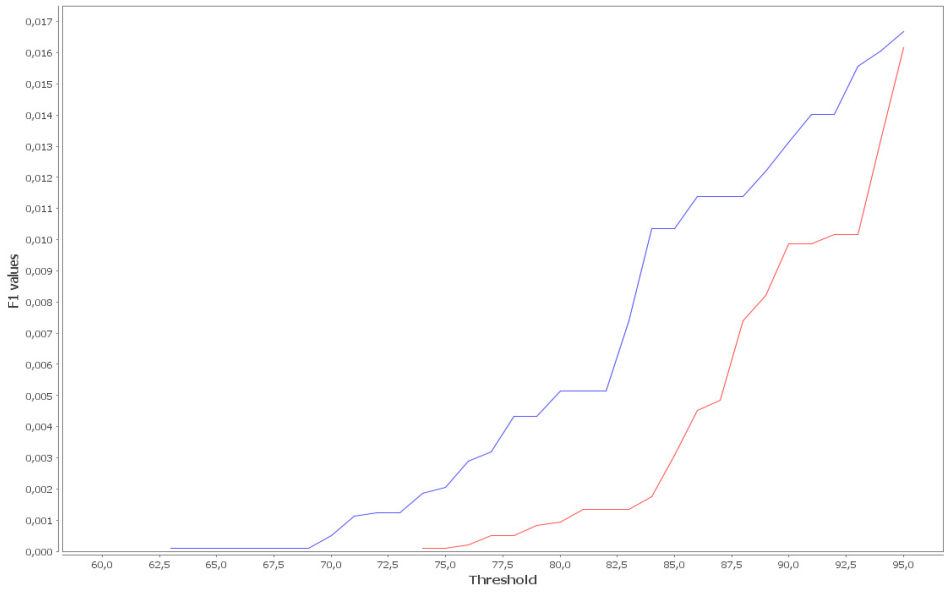


Figure 5.1: F_1 values for different thresholds for primitive PF00069 4-quantile motifs

This graph depicts a set of motifs, derived with different thresholds and assigns their corresponding F_1 values. The red curve describes the behavior of the 4-quantile motifs, while the blue curve is representing the reference motifs, derived from the amino acid sequences of the family alignment.

Figure 5.1 reveals an approximately linear correlation between a chosen threshold and the F_1 measure of the primitive motif finder as a classifier. This indicates that the quality of the classifier is strongly influenced by the threshold that defines a motif. This influence is so enormous that if the threshold is chosen too small, many n-quantile symbols are assigned to the motif, while they are carrying no or only less family-specific profile information. On the other hand if the threshold is chosen too high the generated motif consists of too many wildcard symbols and matches any sequence, which increases the true positive rate artificially, but not significantly. In general it can be stated that the primitive motif finder does not work satisfactory.

5.2 n-Quantile Position Weight Matrices

In section 4.2 the calculation of position weight matrices has been specified to calculate n-quantile PWM's from its general approach as outlined in section 2.4.1. In order to return to the consistent example - the Pfam family PF00062 alignment - the first 40 position weight values for all 4-quantile symbols are listed below as table 5.1:

Table 5.1: Position weight matrix of the PF00062 4-quantile alignment

Table 5.1 lists the log likelihood values assigned to the 4-quantile position weight matrix of the PF000062 alignment.

	Symbol a			
Position	**1**	**2**	**3**	**4**
1	-3.0	-3.0	-3.0	-1.0
2	-3.0	-3.0	-3.0	1.0
3	-3.0	0.0	-1.0	-1.0
4	0.0	0.0	-3.0	-3.0
5	-3.0	-3.0	1.0	-1.0
6	-3.0	1.0	0.0	-3.0
7	0.0	0.0	-3.0	-2.0
8	-3.0	-1.0	0.0	1.0
9	1.0	-3.0	-3.0	-3.0
10	1.0	0.0	-3.0	-3.0
11	-3.0	0.0	0.0	-2.0
12	-3.0	0.0	0.0	0.0
13	1.0	-3.0	-3.0	-3.0
14	-1.0	-2.0	0.0	0.0
15	-3.0	-3.0	0.0	0.0
16	-1.0	0.0	-1.0	-3.0
17	-3.0	-3.0	1.0	-1.0
18	1.0	-3.0	-3.0	-3.0
19	-3.0	-2.0	1.0	0.0
20	-3.0	-2.0	1.0	-2.0
21	-2.0	1.0	-3.0	-3.0
22	-3.0	-1.0	1.0	-2.0
23	-3.0	-3.0	1.0	0.0
24	-1.0	0.0	-2.0	-1.0
25	-3.0	-1.0	1.0	-2.0
26	1.0	-3.0	-3.0	-3.0
27	-2.0	1.0	-2.0	-3.0
28	-3.0	1.0	-1.0	-3.0
29	1.0	-3.0	-3.0	-3.0
30	1.0	-2.0	-3.0	-3.0
31	1.0	-3.0	-3.0	-3.0
32	1.0	-1.0	-3.0	-3.0
33	1.0	-1.0	-3.0	-3.0
34	0.0	-3.0	-1.0	1.0
35	0.0	1.0	-2.0	-3.0
36	-2.0	-1.0	0.0	1.0
37	-3.0	0.0	1.0	-3.0
38	-3.0	-2.0	1.0	-1.0
39	0.0	0.0	-2.0	-3.0
40	-3.0	-2.0	1.0	0.0

Note, that this position weight matrix is transposed. The values assigned to a PWM are often referred to as log-likelihoods. A positive value in the matrix indicates that at this position (for instance the symbol “4” at the second position) a characteristic symbol can be observed. If a symbol occurs at a particular position with a frequency of equal or nearly zero the value assigned to the n-qPWM will be -3.

Instead of using a tabular representation of the PF00062 4-quantile PWM the graph below depicts the log-likelihood values as a profile representation:

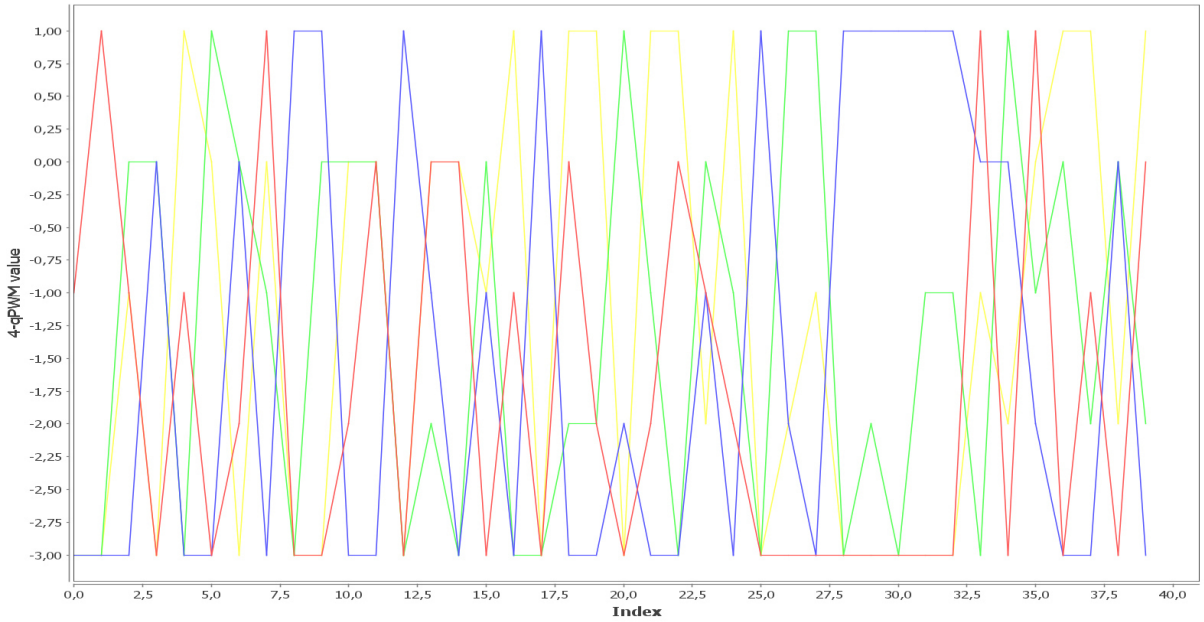


Figure 5.2: 4-quantile PWM log likelihoods of the PF00062 alignment

The coloration of the 4-quantile log likelihood values is analogue to the colorations used in the logo representations with blue as the 4-quantile symbol 1, green standing for 2, yellow for 3 and red for 4.

A comparison of the 4-quantile energy logo shown in section 4.3 as figure 4.1 with the 4-qPWM profile depicted above reveals the parallels of both approaches. Positions which appear as highly conserved in the logo have a relatively high n-qPWM value. If the log-likelihood of a particular 4-quantile symbol occurs as a peak at +1 in this profile representation this symbol can be referred to as highly characteristic for the underlying alignment at this position. In section 6 a short outlook regarding the possible further use of n-quantile position weight matrices and their scores will be given.

5.3 Interpretation of Sequence and Energy Logos

While both approaches the sequence and the energy logo are computed in an analogue way the underlying data basis differs in an important fact: The sequence logo is build up from an alphabet, which contains the 20 canonical amino acids and therefore it carries only the physicochemical information of the distinct amino acids, for instance hydrophobicity or polarity. The energy logo is build up from energy values that are acquiring environmental and local interaction information of the protein structure itself.

Therefore a comparison between both logo types is giving insights into the structural features of a Pfam family alignment which are hidden if the sequence logo is obtained isolated. In order to achieve a higher resolution of the n-quantile symbol distribution 8-quantile logos are used for the comparison:

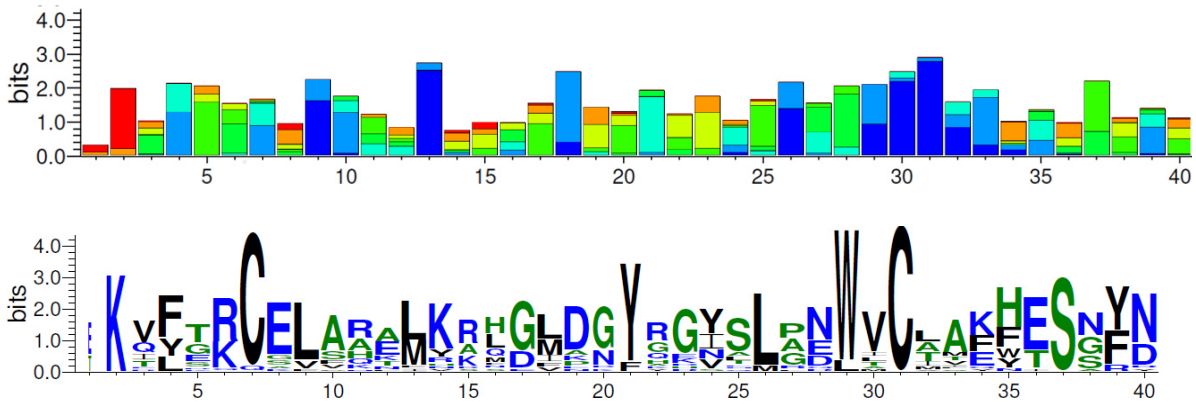


Figure 5.3: Segments of the 8-quantile energy logo and sequence logo of the PF00062 alignment

The first 40 positions of the PF00062 alignment depicted as 8-quantile energy logo and as sequence logo. Note, that the maximum uncertainty of an 8-quantile energy logo is $\log_2(8) = 3$.

The position wise comparison between both logo types provides much room for interpretation. The lysine (K) at the second position is, for instance, realized by high energy values represented as a nearly complete red bar in the energy logo. This behavior of lysine is consistent with its average energy value [76]. More interestingly is the energetic behavior of the cysteines (C) at position 7 and 31.

While the first cysteine at position 7 is realized nearly equally by the second and third quantile in this 8-discretization the second cysteine at position 31 appears as highly conserved in the first quantile. This energetic ambivalence, which cannot be obtained in the sequence logo should be determined in the structure of the observed proteins.

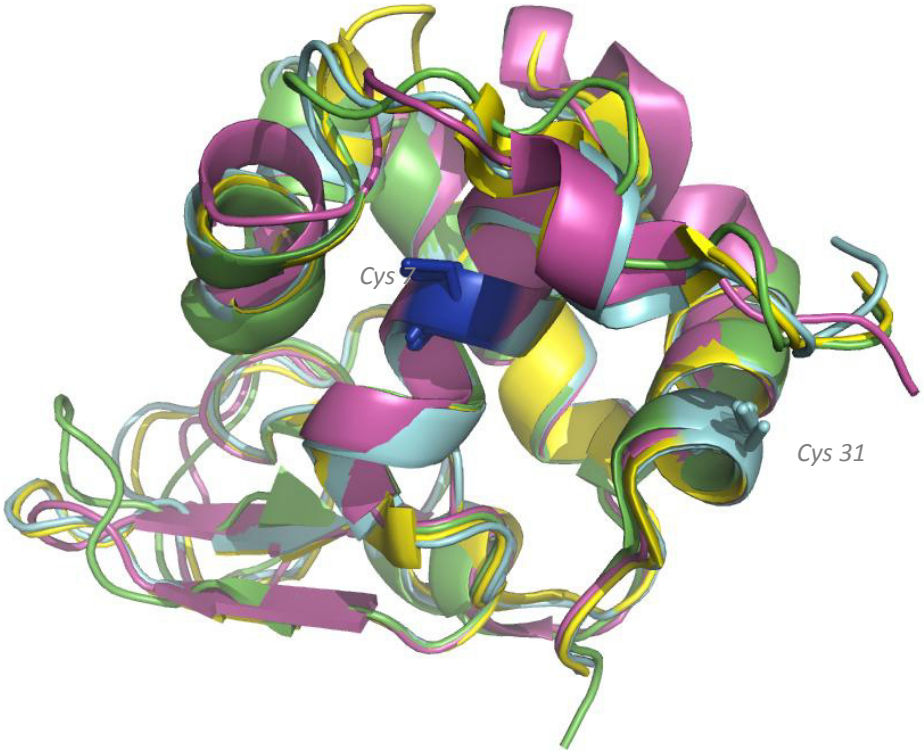


Figure 5.4: Superposition of four structures from PF00062 (1)

Four structures – 1EL1, 1BB6, 1ALC, 2IHL – are picked randomly from the PF00062 family and were aligned with CEalign [67] implemented in PyMOL [68] and their Cys6 (blue) and Cys31 (cyan) residues were labeled to discover their embedment into the structural environment of the proteins.

In figure 5.4 the superposition of four structures from the PF00062 family is depicted and the Cysteines 6 and 31 are labeled. The energy logo in figure 5.3 suggests that these positions are embedded into different structural environments. While the Cys6 is located deep inside an α-helical region the Cys31 is also located in an α-helix, but at the very end and moreover at the surface of the structures facing towards the surrounding solvent. This structural embedment determines the energetic variability, while the observed residues have to be treated equally from a sequential point of view.

However this is an example of sequence conservation that reveals different manifestations on the energetic or, respectively more likely, and the structural level, this circumstance may be also considered other way round. Maybe there are energetic conservations, which are masked behind sequence diversity and thus cannot be detected in the sequence logo representation. The positions 13, where lysine (L) and methionine (M) are observed equally distributed, and especially position 18, which is heavily diverse in the sequence logo, appear as characteristic as realized by the first and the second 8-quantile in the energy logo.

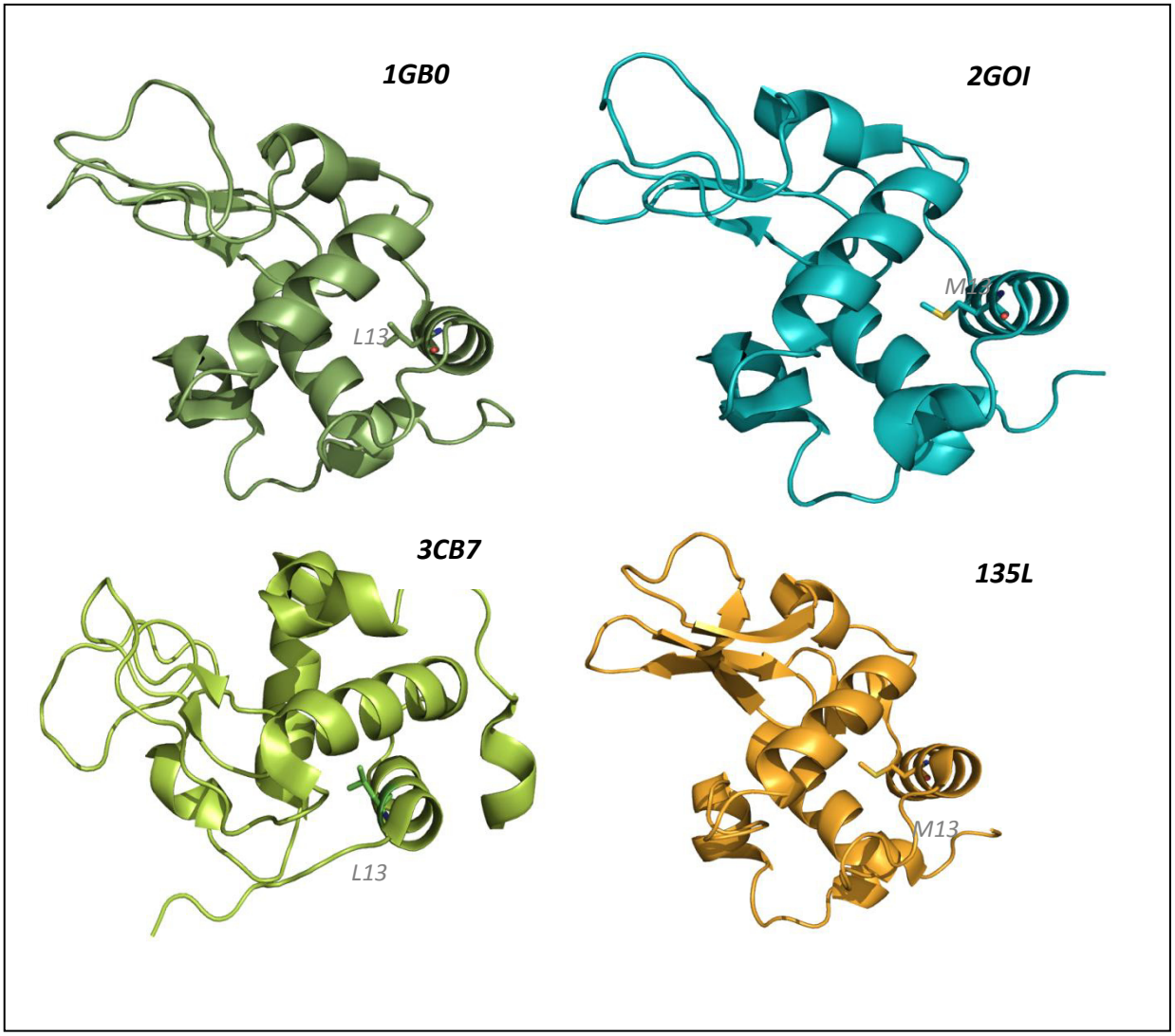


Figure 5.5: Structural embedment of L13/M13 in four structures of PF00062

Four other structures – 1GB0, 2GOI, 3CB7, 135L – are depicted and the lysine/methionine at position 13 in the logo representation shown as figure 5.4 is labeled.

As depicted in figure 5.4 lysine (L) and methionine (M) are equally distributed at the thirteenth position in the sequence logo. The same position seems to be energetic conserved in the energy logo representation, which indicates, that conservation in a general manner does not need to be determined at the sequence level, but on the structural. This means, that a particular position in a structure may occur with high sequence diversity, but is likely to be conserved on the energetic level, as long as it is embedded stable into the whole structure (see figure 5.5).

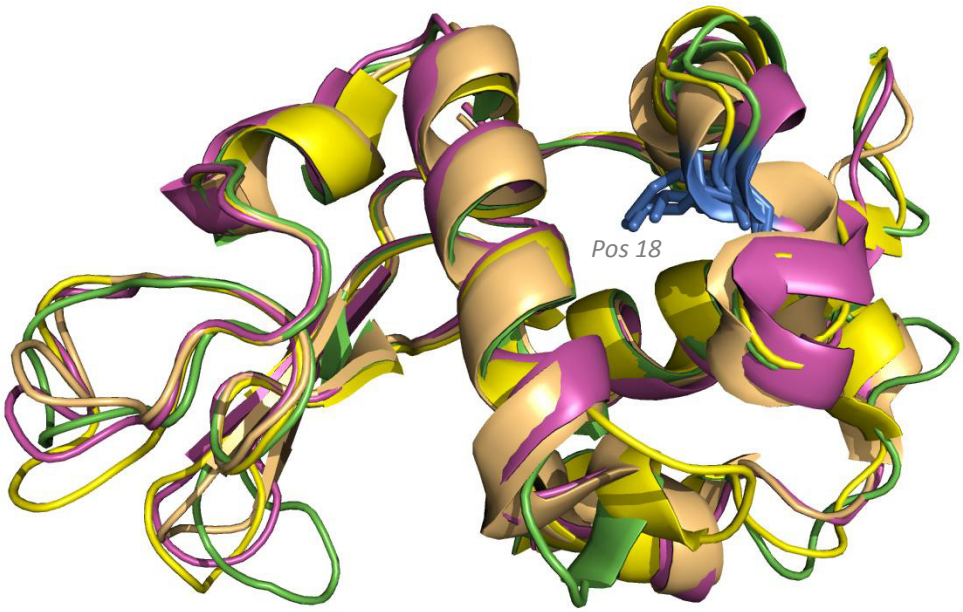


Figure 5.6: Superposition of four structures from PF00062 (2)

Again, four structures – 1EL1, 1ALC, 2IHL, 2G4N – are picked randomly from the PF00062 alignment and were aligned with CEalign in PyMOL. The eighteenth position in the logo representation is labeled blue.

Additional to the fact, that the observed position 18 in the sequence logo shown in figure 5.6 is sequentially highly diverse a look inside the structures as depicted in figure 5.6 reveals, that the residues at this positions a part of different secondary structure elements, which are just barely able to be aligned with the CEalign algorithm. Nevertheless they are well-conserved regarding their energy values due to the fact, that they seem to be embedded into the structures insides.

Considering the cumulative normalized heights mentioned in section 4.4 some regions in the graph assigned with high differences regarding their sequential or energetic conservations can be observed. The cumulative normalized height of the sequence and the 8-quantile energy logo of the Pfam family PF00004 containing 27 alignment entries shall serve as an example.

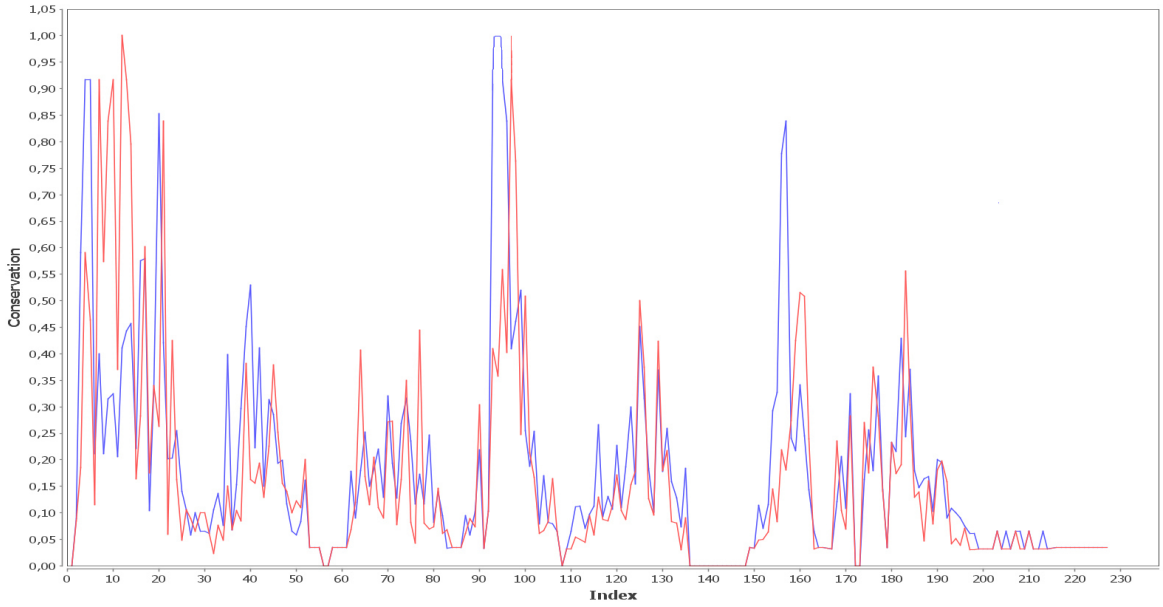


Figure 5.7: Cumulative normalized heights of the sequence and 8-quantile energy logo of the PF00004 family

Depicted are the cumulative normalized heights of both, the sequence and the 8-quantile energy logo of 27 entries in the Pfam family PF00004 alignment. The red curve refers to the sequence logo heights, the blue to those generated from the energy logo.

Especially the region between index 93 and 96, where the energetic conservation is at its maximum, whereas the sequential conservation is, in general lower. Note, that the cumulative normalized height measure does only indicate, that some n-quantile symbols are conserved, but not which in particular. A closer look at this region in the structures of the underlying family alignment entries should reveal in turn the kind of conservation.

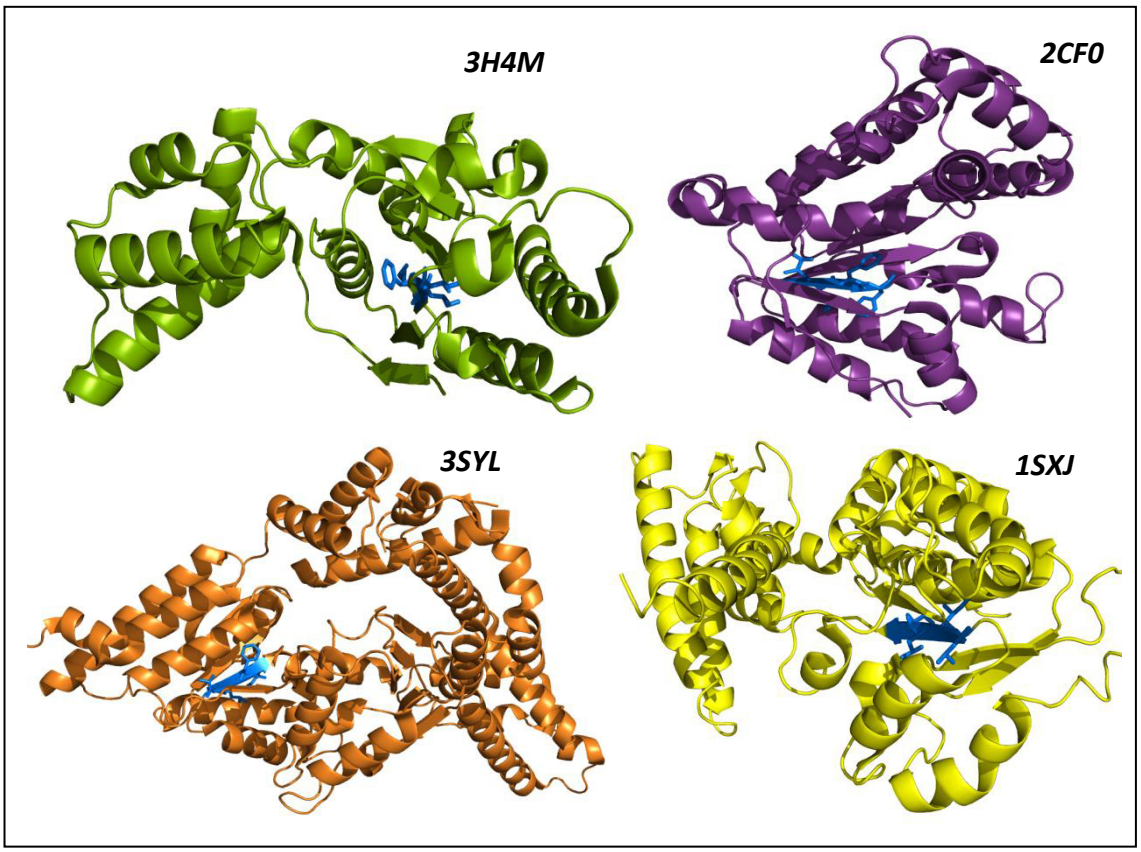


Figure 5.8: Internal β-sheet in four structures of PF00004

Four structures – 3H4M, 2CF0, 3SYL and 1SXJ – were picked randomly from the Pfam family PF00004 alignment and their energetic conserved region mentioned in figure 5.7 is labeled blue and shown as sticks.

In order to be comprehensive and to discover the sequence diversity of this region the corresponding sequence logo is depicted below:

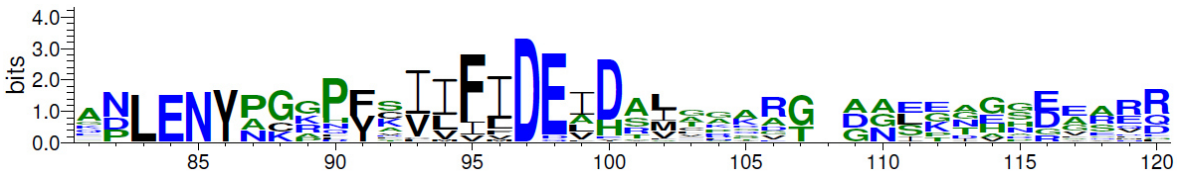


Figure 5.9: Sequence logo segment of the PF00004 alignment

The Classical sequence logo of the 27 amino acid sequences in the Pfam family PF00004 alignment from position 80 to 120.

The sequence logo shows that at the positions 93 – 96 mostly hydrophobic residues like leucine, valine, threonine and phenylalanine are found. These residues are part of the internal β-sheet of the structures and should therefore be assigned with very low energies. By taking a final look at the corresponding 8-quantile energy logo the energetic conservation is obvious to see:

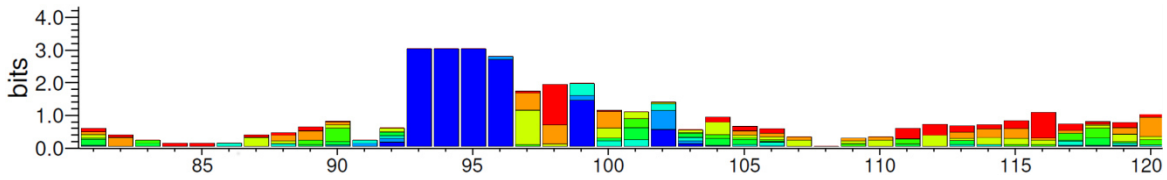


Figure 5.10: 8-quantile energy logo segment of the PF00004 alignment

The 8-quantile energy logo of the 27 8-quantile sequences in the Pfam family PF00004 alignment from position 80 to 120 confirms the assumption.

This family example reveals that energetic conservation is founded in the physico-chemical properties of the residues that build the structure. In other words: The energetic approach generalizes the sequential in this region, so that shared physic-chemical properties among a set of diverse residues are translated into energetical conservation.

Another fact that deserves attention is the question: Is there any correlation between sequential and energetic conservation? Therefore the sequence and energy logos of all families in the dataset were generated and the spearman correlation coefficient of the cumulative normalized height (CNH) curve of the sequence logo and of the corresponding n-quantile energy logo was calculated.

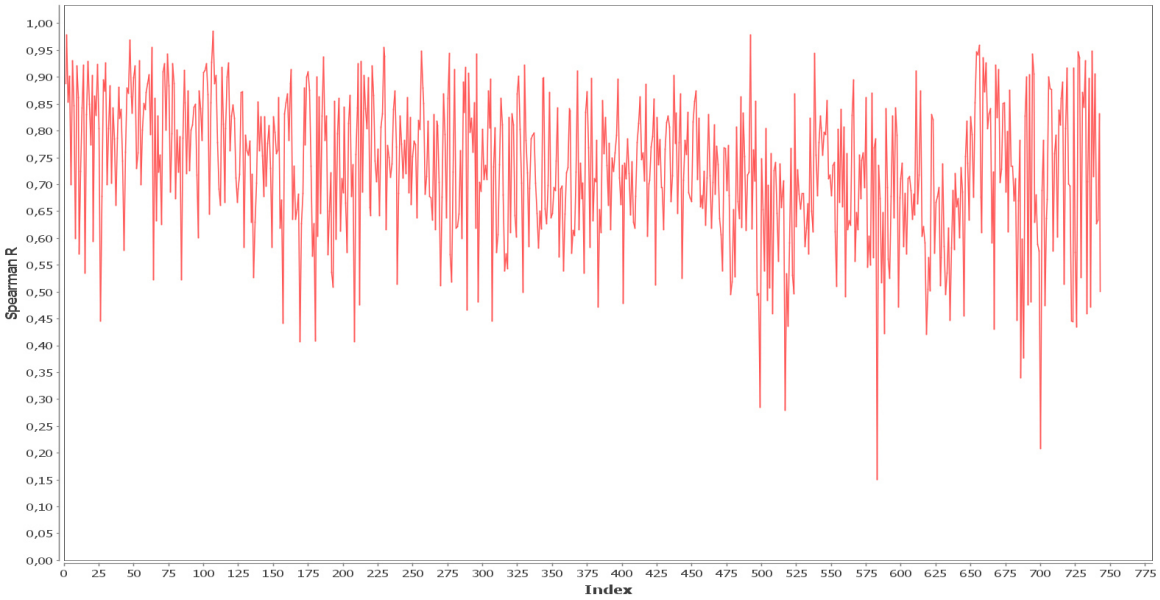


Figure 5.11: Spearman correlation coefficients between the CNH-curves of 743 sequence and 4-quantile logos

Here a datapoint is the spearman correlation coefficient between the cumulative normalized height curve of a sequence and a 4-quantile energy logo of one family in the whole family dataset.

The plotting of 743 spearman correlation coefficients between the cumulative normalized height curves of a family-specific sequence and 4-quantile energy logo reveals that some logos of a particular family correlate well with values bigger than 0,9, whereas some do not and are assigned with values of 0,5 or even below.

At different n-discretizations the behavior of the correlation of the CNH-curves of both logo types does not change.

Table 5.2: Mean spearman coefficients of CNH-curves of sequence and energy logos at different values of n

Averaged spearman R's of CNH -curves of sequence and energy logos at different n's. Every mean value represents 743 value pairs.

Discretization	Mean Spearman R
n = 4	0.728
n = 6	0.746
n = 8	0.748
n = 10	0.744
n = 12	0.733

6 Outlook

While a large part of the possible work with the depeloped energy logos is presented in this work, the n-quantile position weight matrix (*n*-qPWM) approach may promise more applications in the future. During the studies of the *n*-qPWM's the score-assignment approach has proved itself as not that applicable in the current form. Therefore another approach – the *n*-qPWM structure classifier – shall be suggested to answer the principle possible question: To which Pfam family does a given functionally unanotated protein structure S belong? The following pseudocode lists the different steps of the idea:

I. Generate the energy profile E_S to S
II. Discretizise E_S into n-quantile sequences, where n is arbitrary
III. For (Pfam family energy profile set PFEPS from family dataset)
{
Align E_S to PFEPS via MEPAL
Calculate n-qPWM from all n-quantile sequences in PFEPS
Calculate n-qPWM Score for E_S
}
IV. Determine the best score

Finally, the best, which would be the highest, score of E_S to a Pfam family energy profile set PFEPS indicates the most propable candidate family for E_S. To ensure that the results that are as reliable as possible a Pfam family energy profile set needs to contain a sufficient amount of underlying energy profile data. Additionally, the score calculation between a query E_S and an n-qPWM derived from a set of Pfam family energy profiles could be handled via dynamic programming instead of a naive approach.

Literature on calculating log-likelihood values in position weight matrices suggests that the inclusion of substitution values for amino acids leads to much more accurate results. [55] Therefore the use of a substitution model for discretizised energy values as given by M. Langer [69] may lead to high quality position weight matrix or even the construction of well-performing profile-hidden markov models to define family specific n-quantile patterns.

Another important circumstance, which is yet to be done, regards the further development of the energy logo. The assumption that the distinct n-quantile symbols are distributed equally among a sufficient big set of energy data is deniable as well as it's for amino acids plotted in a sequence logo. Although the particular background frequencies of the different n-quantile symbols were computed during the studies with n-quantile position weight matrices it remains unclear whether the underlying dataset is big and purposeful enough to serve as a representative set nor how the relative entropy of a particular n-quantile symbol in an alignment should be calculated.

7 Summary

This work is focused on the question how pattern or motif information could be extracted out of a set of quantilized family energy profiles.

The approach to include methods inspired by information theory, such as the development of the energy logo to obtain a well-curated set of Pfam families delivered valuable results. Although no concrete motifs that could characterize a particular family were generated due to the high complexity of this topic, this work did open the entrance to further work on finding energetic relevant motifs in families or even fold-classes.

The obtainment of conservation at both, sequential and energetic level has revealed that the classic term of conservation might be expanded. The development and discussion of the quantile energy logos has proven that such logos are able to reveal structural conserved sites in proteins. As shown in section 5 the understanding of conservation cannot be treated at the sequence level only, as it's a more complex rule in biological systems, which extends over the abstraction levels of sequence, structure and energy. Therefore we might assume, that protein structures are not build up out of amino acids or secondary structure elements in the classical understanding, but are more likely to consist of sensible composed structurally and energetic instances.

Bibliography

[1] E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin, Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments, PROTEINS: Structure, Function, and Genetics 28:405–420, 1997.

[2] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton, CATH — a hierarchic classification of protein domain structures, Structure Vol 5 No 8, 1997.

[3] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Research, 2014, Vol. 42, Database issue, 2013.

[4] F. Heinke, S. Schildbach, D. Stockmann, D. Labudde, eProS—a database and toolbox for investigating protein sequence–structure–function relationships through energy profiles, Nucleic Acids Research, Vol. 41, 2013.

[5] E. G. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, WebLogo: A Sequence Logo Generator, Genome Research, 14:1188-1190, 2004.

[6] G. D. Stormo, T. D. Schneider, L. Gold, A. Ehrenfeucht, Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli, Nucleic Acids Research 10 (9): 2997–3011, 1982.

[7] E. Y. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl, F. U. Hartl, Molecular chaperone functions in protein folding and proteostasis, Annual Review of Biochemistry, 82(1):323–355, ISSN 1545-4509. doi: 10.1146/annurev-biochem-060208-092442, 2013.

[8] C. Levinthal, How to Fold Graciously, 1969.

[9] P. Y. Chou, and G. D. Fasman, Prediction of protein conformation, Biochemistry, Jan; 13(2), pp 222–245, 1974.

[10] J. Garnier, J. F. Gibrat , B. Robson, GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence, Methods in Enzymology, Vol. 266, 1996.

[11] Y. Zhang, I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, vol 9, 40, 2008.

[12] J. Söding, A. Biegert, A. N. Lupas, The HHpred interactive server for protein homology detection and structure prediction, Nucleic Acids Research 33((Web Server issue)): W244–248, 2005.

[13] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling, Bioinformatics, 22,195-201, 2006.

[14] J. C. Kendrew, et al., A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis, Nature 181 (4610): 662–6, 1958.

[15] Protein Data Bank, URL: <http://www.rcsb.org/pdb>, last available on 21th, August, 2014

[16] American Crystallographic Association, URL: <http://www.amercrystalassn.org/content/images/History/XrayDiff3.gif>, last available on 21th, August, 2014.

[17] Oxford Journals, URL: <http://bioinformatics.oxfordjournals.org/content/23/21/2851/F2.expansion.html>, last available on 21th, August, 2014.

[18] Nature, URL: <http://www.nature.com/nature/journal/v509/n7501/fig_tab/nature13167_SF2.html>, last available on 21th, August, 2014.

[19] M. J. Sippl, Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures, J. Computer-Aided Mol. Des., 7(4):473–501, Aug 1993.

[20] H. J. Dyson, M. Rance, R. A. Houghton, R. A. Lerner, P. E. Wright, J. Mol. Biol., 201 161, 1988.

[21] C. B. Anfinsen, Principles that govern the folding of protein chains, Science, vol. 181, no. 4096, pp. 223–230, 1973.

[22] S. Tanaka and H. A. Scheraga, Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins, Macromolecules, vol. 9, no. 6, pp. 945–950, 1976.

[23] D. H.Wertz and H. A. Scheraga, Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule, Macromolecules, vol. 11, no. 1, pp. 9–15, 1978.

[24] D. Labudde, F. Heinke, Membrane Stability Analysis by Means of Protein Energy Profiles in Case of Nephrogenic Diabetes Insipidus, Computational and Mathematical Methods in Medicine, 2012.

[25] F. Dressel, A. Marsico, A. Tuukkanen, M. Schroeder, and D. Labudde, Understanding of SMFS barriers by means of energy profiles, in Proceedings of the German Conference on Bioinformatics, pp. 90–99, 2007.

[26] Adaptiert nach F. Heinke, 2012.

[27] Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

[28] Energy Profile Suite, Bioinformatics Group Mittweida, URL: <http://bioservices.hs-mittweida.de/Epros/>, last available on 21th, August, 2014.

[29] T. Bailey, C. Elkan, Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization, Department of Computer Science and Engineering University of California, San Diego, 1997.

[30] C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, J. Wootton, Detecting Subtle Sequence Signals: A Gibbs Sampling Stragety for Multiple Alignment, Science, New Series, Volume 262, Issue 5131, 208-214, 1993.

[31] J. Huang, D. Brutlag, The EMOTIF database, Nucleic Acids Research, Vol 29, No. 1, 2001.

[32] S. Pietrokovski, J. G. Henikoff and S. Henikoff, The Blocks database - a system for protein Classification, Nucleic Acids Research, Vol. 24, No. 1, 1995.

[33] T. K. Attwood, H. Avison, M. E. Beck, M. Bewley, A. J. Bleasby, F. Brewster, P. Cooper, K. Degtyarenko, A. J. Geddes, D. R. Flower, M. P. Kelly, S. Lott, K. M. Measures, D. J. Parry-Smith,‖ D. N. Perkins, P. Scordis, D. Scott, C. Worledge, The PRINTS Database of Protein Fingerprints: A Novel Information Resource for Computational Molecular Biology, J. Chem. Inf. Comput. Sci., 37 (3), pp 417–424, 1997.

[34] X. Chen, L. Guo, Z. Fan, T. Jiang, Learning Position Weight Matrices From Sequence And Expression Data, 2007.

[35] J. Li, Y. Zhang, W. Qin, Y. Guo, L. Yu, X. Pu, M. Li, J. Sun, Using the improved position specific scoring matrix and ensemble learning method to predict drug-binding residues from protein sequences, Natural Science, Vol. 4, No. 5, 304-312, 2012.

[36] I. Piedade, M. E. Tang, O. Elemento, DISPARE: DIScriminative PAttern REfinement for Position Weight Matrices, BMC Bioinformatics, 10:388, 2009.

[37] F. Heinke, Energieprofilbasierende Analysemethoden von Proteinfamilien, Mittweida, 2010.

[38] Krogh, A., Brown, M., Mian, S., Sjölander, K. and Haussler, D., Hidden Markov models in computational biology, Applications to protein modeling. J. Mol. Biol., 235, 1501–1531, 1994.

[39] S. R. Eddy, Profile hidden Markov models, Bioinformatics, 14, 755–763, 1998.

[40] Pfam Protein Database, URL: <http://pfam.xfam.org>, last available on 21th, August, 2014.

[41] S. R. Eddy, A new generation of homology search tools based on probabilistic inference, Genome Informatician, 23, 205–211, 2009.

[42] S. R. Eddy, Accelerated profile HMM searches. PLoS Comput. Biol., 7, e1002195, 2011.

[43] UniProt Protein Knowledgebase, URL: <http://www.uniprot.org>, last available on 21th, August, 2014.

[44] PROSITE, Database of protein domains, families and functional sites, URL: <http://prosite.expasy.org>, last available on 21th, August, 2014.

[45] J. P. Overington, Comparison of three-dimensional structures of homologous proteins. Structural Biology, 2:394–401, 1992.

[46] E. L. L. Sonnhammer, D. Kahn, Modular arrangement of proteins as inferred from analysis of homology, Protein Science, 3:482–492, 1994.

[47] National Center for Biotechnology Information, URL: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>, last available on 21th, August, 2014.

[48] E. L. L. Sonnhammer, R. Durbin, A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis, Gene 167:GC1–10, 1996.

[49] J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22:4673–4680, 1994.

[50] D. G. Higgins, A. J. Bleasby, R. Fuchs, CLUSTAL V: Improved software for multiple sequence alignment. Comput. Appl. Biosci. 8:189–191, 1992.

[51] G. D. Stormo, DNA binding sites: representation and discovery, Bioinformatics 16 (1): 16–23, 2000.

[52] S. Sinha, On counting position weight matrix matches in a sequence, with application to discriminative motif finding, Bioinformatics 22 (14): e454–e463, 2006.

[53] X. Xia, Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction, Scientifica 2012: 1–15, 2012.

[54] S. Henikoff J. G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA Vol. 89, pp. 10915-10919, 1992.

[55] M. Zvelebil, J. O. Baum, Understanding Bioinformativs, pp. 169-173, 2008.

[56] R. Guigo, An Introduction to Position Specific Scoring Matrices, URL: <http://bioinformatica.upf.edu/T12/MakeProfile.html>, last available on 21th, August, 2014.

[57] T. D. Schneider, R. M. Stephens, Sequence logos: A new way to display consensus sequences. Nucleic Acids Res. 18: 6097–6100, 1990.

[58] C. E. Shannon, Bell System Tech. J. 27, 379-423, 623-656, 1948.

[59] J. R. Pierce, An Introduction to Information Theory: Symbols, Signals and Noise, Dover Publications, Inc., New York second edition, 1980.

[60] A. P. Bercoff, T. R. Bürglin, J. Koch, LogoBar – Visualizing Protein Sequence Logos with Gaps, Bioinformatics, Jan 1;22(1):112-4, 2006.

[61] G. P. Basharin, Theory Probability Appl., 4, 333-336, 1959.

[62] G. A. Miller, Note on the bias of information estimates, Information Theory in Psychology, II-B, 95-100, 1955.

[63] T. D. Schneider, G. D. Stormo, L. Gold, A. Ehrenfeucht, Information content of binding sites on nucleotide sequences. J.Mol. Biol. 188: 415–431, 1986.

[64] S. B. Needleman, C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology, 48, S. 443–453, 1970.

[65] F. Heinke, M. Langer, Personal communication, Bioinforamtics Group Mittweida (BiGM) 2014.

[66] URL: http://bioservices.hs-mittweida.de/Epros/Index?page=stats, 2014.

[67] I. N. Shindyalov, P. E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, Protein Eng. Sep; 11(9):739-47, PMID: 9796821, 1998.

[68] The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC, 2014.

[69] M. Langer, Entwicklung molekular-phylogenetischer Methoden auf Grundlage von Aminosäurerest-Pseudopotentialen, Hochschule Mittweida (FH), 2014.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, 21. August 2014