Empirical Protein Potentials From Delauney Tesselation

Diplomarbeit

zur Erlangung des akademischen Grades

Magister rerum naturalium

an der Formal- und Naturwissenschaftlichen Fakultät
der Universität Wien

Vorgelegt von

Günther Weberndorfer

im Jänner 1999

Für meinen Vater, der diesen Tag leider nicht mehr erleben durfte.

Vorwort

Diese Arbeit ist in der Zeit von März bis Dezember 1998 am Institut für theoretische Chemie und Strahlenchemie der Universität Wien entstanden, und wurde von Peter Stadler und Ivo Hofacker betreut.

Bei Ihnen möchte ich mich an dieser Stelle besonders bedanken. Peter Stadler unterstüzte mich durch seine wissenschaftliche Leitung, sowie sein überwältigendes Wissen, und ohne Ivo Hofacker hätte vieles sehr sehr viel länger gedauert. Danke — es war schön mit Euch zu arbeiten.

Prof.Peter Schuster danke ich für die Inspiration, meine Diplomarbeit am Institut für theoretische Biochemie zu machen, wie auch für die freundliche Aufnahme in seine Arbeitsgruppe.

Meinen Kollegen vom TBI, Ronke Babajide, Martin Fekete, Christoph Flamm, der mir unglaublich oft geholfen hat, Thomas Griesmacher, Stephan Kopp, Bärbel Krakhofer, Stefan Müller, Susanne Rauscher, Alexander Renner, der mich sehr viel gelehrt hat, Roman Stocsits und Andreas Wernitznig, danke ich für die beste Zeit, die ich während meines gesammten Studiums hatte.

Besonderen Dank schulde ich meinen geliebten Eltern und Großeltern, die mich während meiner gesammten Studiendauer aufs beste unterstüzt haben, und mir diesen Weg ermöglicht haben. Ohne die Unterstützung meiner Familie hätte ich nicht studieren können.

Am allermeisten aber verdient meine Verlobte Sylvie Meraner, meinen Dank. Sie mußte während all der harten Studienjahre meine Höhen und Tiefen ertragen. Erst sie hat mir immer wieder die Kraft gegeben weiter zu machen.

Inhaltsverzeichnis

Zusammenfassung

Eine überwältigende und immer schneller wachsende Flut an Sequenzinformation aus groß angelegten Sequenzierexperimenten erschwert immer mehr den Blick auf relevante Daten. Man kann zwar DNA Sequenzen leicht in die entsprechende Proteinsequenz „translatieren“, jedoch ist diese ohne die dazugehörige $3d$ Struktur oft nutzlos. Ein detailliertes „mapping“des astronomisch großen Proteinsequenzraumes wäre völlig aussichtslos, wenn nicht die Zahl der stabilen Strukturen sehr begrenzt wäre. Diese Kartierung wäre von großem Nutzen für evolutionäre Studien und das *de novo* Design von Proteinen. Babajide *et al.* konnten zeigen, daß empirische Potentiale zur Erforschung des Proteinraums mittels inverser Faltung geeignet sind. Hierbei wird die Kompatibilität einer Struktur mit einer gegebenen Sequenz ermittelt. Die zugrunde liegenden Annahmen sind, daß das Protein im energetischen Grundzustand vorliegt, und das inverse Boltzmann Gesetz gilt. Da die Grundzustandsenergie nicht bekannt ist, d.h. das Faltungsproblem nicht gelöst ist, muß eine Energieskala, z-score genannt, eingeführt werden, um Vergleiche anstellen zu können.

Empirische Potentiale, die aus einer Datenbank strukturelle Informationen „extrahieren“, unterscheiden sich meist in der Definition der berücksichtigten Kontakte. Alexander Tropsha konnte die Willkür eines gewählten Abstandes umgehen, indem er Methoden der statistischen Geometrie einführte. Die Proteinkette wird hierzu durch die C^{α}-Atome dargestellt. Die dadurch definierte Menge von Punkten im Raum wird der Delauney Tessellation unterzogen.

Das Ergebnis ist ein Agglomerat dicht gepackter, unregelmäßiger Tetraeder, mit einer Aminosäure an jeder Ecke. Diese Beschreibung einer nächsten Nachbarschaft wird verwendet, um die Statistik der Wechselwirkung in einer Untermenge der pdb-Datenbank zu ermitteln. Dies ermöglicht die Berechnung der Wahrscheinlichkeit, diesen bestimmten Kontakt vorzufinden.

Unglücklicherweise zeigten inverse Faltungsexperimente, die mit diesem Potential durchgeführt worden sind, Inkonsistenzen mit Daten aus anderen Potentialen (wie zum Beispiel `PROSA`). Das führte zu der Idee, das Tesselations Potential zu erweitern: Ein Oberflächenterm sollte die spezielle Rolle der lösungsmittelexponierten Reste berücksichtigen. Außerdem wollte man dazu übergehen, das Backbone durch C^{β}-Atome darzustellen, da diese räumlich in Richtung der Aminosäurereste zeigen. Es war auch nötig ein Filterkriterium anzuwenden, um unwahrscheinliche Kontakte zu entfernen, die aus dem Tessellationsalgorithmus der konvexen Hülle stammen.

Diese Arbeit beschreibt die erfolgreiche Implementierung eines Tessellations Potentials und dessen Anwendung in Simulationen von inversen Proteinfaltungen. Die Computerexperimente zeigen eindrucksvoll eine deutliche Verbesserung gegenüber den Originaldaten von Alexander Tropsha. Auf Grund der effizienten Implementierung des Kalibrierungsvorgangs ist es nun einfach, weitere Zusatzterme einzuführen.

Abstract

The overwhelming and fast growing amount of known sequences from large scale sequencing projects hardly leaves a view to the relevant data. It is easy to translate the DNA sequence to the corresponding protein chain, but by the time impossible to gain access to the structure from this information level, because the amino acid sequence separate from its $3d$ context is often meaningless. Nevertheless sequence homologies and alignement studies are useful tools. A detailed mapping of the hyper-astronomic sequence space of proteins would be a hopeless task, if the number of distinct stable folds would not be restricted. Such a map would be extremely useful for evolutionary studies as well as protein *de novo* design.

Babajide and co-workers revealed that knowledge based potential are suitable means to perform an analysis of protein space, targeting inverse folding. This approach tries to determine the compatibility of a given structure with a chosen sequence. The basic assumptions of knowledge based potentials are that proteins exist in the energetic ground state and the inverse Boltzmann law is valid. Since the ground state is not known (i.e. the folding problem is not solved), an energy scale, the so called z-score, must be introduced for comparison.

Empirical potentials as extracted form databases of known structures vary mainly in the definition of residue interaction. Avoiding the arbitrariness of a binned distance, Alexander Tropsha introduced a "statistical geometry" approach, in which the polypeptide chain is un-ambiguously partitioned by means of Delauney tessellation. The result is a cluster of tightly packed, irregular tetrahedra having an amino acid at each corner. This description of contact is used to extract a log-likelihood quantity for various types of interactions from a subset of the pdb database.

Unfortunately the inverse-folding experiments performed with this kind of potential showed a severe inconsistence with other potentials such as `PROSA`. This led to the idea of extending the original potential: A special surface term should be introduced, paying respect to the special role of solvent exposed residues. Furthermore the chain representation should be extended to C^{β} atoms because they point towards the residue. It was also necessary to introduce a filter criterion to the calibration method for removing improper contacts as produced by the convex hull.

This work describes a successful implementation of a tessellation potential and a few applications to inverse folding computer experiments. A distinct improvement compared to the original parameters could be obtained and directions for further improvements of the discrimination power are pointed out.

1 Introduction

Since the early 19^{th} century it is known that, according to Lamarck, "*each science has to have its philosophy. Only then real progress is possible.*"[38]. Projecting this to modern biology of the late 20^{th} century, a new meaning is gained by this sentence. The fast growing wealth of data from cloning and sequencing projects hardly leaves space in molecular biology for finding theories and organizing the empirical knowledge, the *philosophy* of science as we understand it today is considered to be secondary.

The striking breakthroughs in molecular biology were mostly brought by the knowledge of the participating structures of a biochemical process. In particular for proteins, being the "genetic executable" the $3d$ structure is an avenue to understand function on the molecular level. It is almost trivial to translate the DNA sequence, if known, to the corresponding protein, but a protein sequence contains little meaning unless in the context of its spatial distribution and interaction. Structures are the key to understanding function, though the exact determination of dynamics is still one step after the folding problem.

Modeling structures of biomolecules ahead of experiments is still a demanding challenge, up to now the answer to the folding problem can only be given for the comparable simple logic of nucleic acid secondary structures [32, 62, 71, 72]. For proteins a mapping of sequence to structures is still out of sight, though intensive research opens the view into this world of complexity.

In contrast to nucleic acids, where the main part of the folding energy derives from basepair stacking, the driving force for protein folding are the more or less unspecific hydrophobic interaction. These hydrophobic contributions are hardly characterized or measured. It is widely assumed that the native structure of a protein represents the global minimum of its the energy function $W(S)$. The number of terms contributing to the energy function is enormous and depend on the amino acid composition as well as the natural environment (pH, temperature, ionic strength, solvent type etc.), but if the function would be known, the sequence S could in principle be assigned to a fold $\psi(S)$. So it would be very favorable to gain access to structures form sequence databases using the concept of "data-mining" as introduced in computer science, and to avoid a detailed determination of all parameters. Extracting structural information from sequence databases, in other words solving the folding problem, is still out of sight though much work is done on this topic.

Some approaches targeting protein secondary structure [16, 45, 61] brought reasonable success, for instance Sander *et al.* [48]) report an accuracy of some 70% in predicting the structural elements from sequence. But it has to be considered carefully that assigning the random coil as local structure is counted as predictional success. This is not very meaningful since usually two different conformations

in random coil have a large root-mean-square error when superimposed, and can not be considered to be similar. Tertiary structure prediction is targeted by potential energy based analysis (e.g. molecular dynamics), lattice simulations of protein folds, and knowledge based approaches. However, accurate results are rare and mostly restricted to cases where homologous structures are known [47]. The nature of this optimization problem makes all kinds of calculation extremely costly in computation time. Though a lot of rules for protein folding have been discovered, at the moment the only straightforward way to get $3d$-structures are NMR spectroscopy and X-Ray crystallography.

The size of the sequence space [60] for proteins is enormous, since it grows exponentially with the chain length (20^n sequences for chain length n). On the other hand, the amount of stable folds seems to be small [33], therefore it makes sense to ask, how sequences adapting a similar fold are distributed in sequence space. The sequences folding into a given structure ψ form the *neutral* set $S(\psi)$ of this structure. This gives rise to a seizable problem — the question of threading a sequence to a known native structure, this is *inverse folding*. If all possible structures were known, the folding problem would reduce to the compatibility of a sequence with a given structure. Detailed knowledge about the topology of the neutral sets in sequence space containing is important to answer questions arising from protein evolution and *de novo* protein and drug design as targeted by industrial applications.

Again the situation for nucleic acids is much simpler from the predictional point of view, and some very unexpected results were obtained by exploring the sequence-structure maps for RNA. It could be shown, that expanded neutral networks percolate the entire sequence space, and in some parts sequences of different structures come very close to each other [27, 28, 36].

Babajide *et al.* [3, 2] showed that knowledge based potentials are suitable tools to investigate neutrality in protein space. The idea of using a database of known structures to obtain information about spatial distribution of residues is quiet old (see Blundell [7]), and the idea of using a statistical mechanical interpretation arose in the early 1990s. The models differ mainly in the definition of interaction and unfortunately yield oftentimes diverging results.

The methods of computational geometry may help to avoid this inconsistence by applying objective criteria to neighborhood definition. Bernal [5] proposed in the late 1950ies to characterize disordered systems by means of irregular polyhedra as obtained by a specific tessellation in three dimensional space. Representing an amino acid chain by its C^α atoms yields a set of points in $3d$-space, uniquely describing the backbone of the protein. Applying the *Delauney Tessellation* generates a tightly packed cluster of space filling irregular tetrahedra (called *Delauney simplices*) with four C^α atoms forming the corners.

This approach has successfully been introduced by Alexander Tropsha *et al.* [68, 54, 69] in 1996 but employing his original parameters led to the observation, that the energies (as represented by z-scores) showed some inconsistency with potentials derived from other approaches such as Sippl's `PROSA` potential [55, 13, 56] and hence we decided to extend the original potential by introducing a surface term and to change the representation of the backbone from C^{α} to C^{β} atoms. The fact that there was no appropriate tool to calibrate the Tropsha potential made a design of such a program unavoidable.

Organization of this Work

The first part of this work gives a brief overview of potential functions used in recent computer experiments comparing molecular mechanical approaches and then presenting various knowledge based potentials. The main focus will the be on the development of tools to calibrate and extend the Tropsha-like potential function. Finally details of the algorithms are described and a brief manual is included. Then we present first results obtained from the improved potentials.

2 Theoretical Background

2.1 Molecular Force-Fields versus Knowledge-Based Potentials

The energy of a macromolecular system is a function of the conformational variables (e.g. Cartesian coordinates) plus its interaction energy with the surrounding solvent. The derivation of the energy from the conformational variables gives the force field of the molecule. The term potential in this context is a synonym to energy function. Generally we assume that a protein sequences $S = (s_1, \ldots, s_n)$ of n amino acids

$$s_i \in \{\mathbf{A}, \mathbf{C}, \mathbf{I}, \mathbf{L}, \mathbf{M}, \mathbf{F}, \mathbf{W}, \mathbf{Y}, \mathbf{V}, \mathbf{R}, \mathbf{N}, \mathbf{D}, \mathbf{E}, \mathbf{Q}, \mathbf{G}, \mathbf{H}, \mathbf{K}, \mathbf{P}, \mathbf{S}, \mathbf{T}\}$$

is related with its structure ψ as represented by the coordinates $x_s = (x_1, \ldots, w_n)$ via the potential function $V(S, \psi)$:

$$x_s = \text{argmin}_x V(S, \psi)$$

The design of molecular force fields allows at least two different approaches:

On the one hand semi-empirical approaches consider macromolecular systems as a summation of the forces observed for monomers. The force fields are obtained from quantum mechanical calculations, and data from thermodynamic or spectroscopic measurements on small molecules.

On the other hand knowledge-based potentials are based on the assumption that force fields of macromolecules are of immense complexity and the only reliable source of information are macromolecular molecules themselves. So empirical or knowledge-based potentials try to extract information from databases of macromolecular structures.

2.2 Molecular Mechanics Force Fields

The "mechanical" molecular model was developed out of the need to describe molecular structures and properties in as practical a manner as possible. Quantum chemical calculation are highly accurate, but the computational effort is immense and at the moment it is unthinkable to solve the Schrödinger equation for macromolecular systems. Therefore, a classical approach was chosen to calculate atomic structures. The following assumptions are made:

- Nuclei and electrons are lumped into atom-like particles according to the Born-Oppenheimer approximation of the Schrödinger equation.

- The atom-like particles are spherical (radii obtained from measurements or theory) and have a net charge (obtained from theory).
- Atoms are considered as balls, bonds are based on springs, whereby classical potentials come to use.
- Interactions must be preassigned to specific sets of atoms.
- Interactions determine the spatial distribution of atom-like particles and their energies.

The object of molecular mechanics is to predict the energy associated with a given conformation of a molecule. However, molecular mechanic energies have no meaning as absolute quantities. Only energy differences between two conformations of the same molecule are meaningful. A simple molecular mechanic energy equation is given by:

$$E_{tot} = E_{stretch} + E_{bend} + E_{tors} + E_{non-bonding}$$

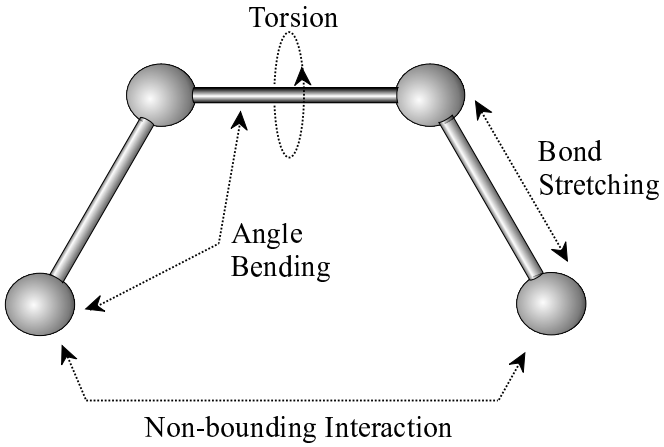


Figure 1: Energies used by molecular mechanic force fields

These terms together with the parameters required to describe the behavior of different kinds of atoms and bonds, is called a force-field. Many different kinds of force-fields have been developed over the years. Some include additional energy terms that describe other kinds of deformations. Some force-fields account for coupling between bending and stretching in adjacent bonds in order to improve the accuracy of the mechanical model. The constants (force constants, equilibrium lengths) can be either measured by spectroscopy or calculated by quantum mechanical means.

The energy terms in detail are:

Stretching Energy:
Occurs whenever a bond is deformed (stretched or compressed), and is described by an equation based on Hooke's law for springs.

$$E_{stretch} = \sum k_b(r - r_0)^2$$

whereby k_b is the force constant, r is the actual bond length and r_0 the equilibrium length. This parabolic approximation fails as the bond is stretched toward the point of dissociation.

Bending Energy:
Energy increases if the equilibrium bond angles are bent. Again the approximation is harmonic and uses Hooke's law.

$$E_{bend} = \sum k_\theta(\theta - \theta_0)^2$$

k_θ controls the stiffness of the angel, θ is the actual bond angle, θ_0 the equilibrium angle. The force constants have to be estimated for each triple of atoms (e.g. C-C-C, C-C-O, C-C-H)

Torsion Energy:
Intra-molecular rotations (around torsions or dihedrals) require energy as well:

$$E_{torsion} = \sum A[(1 + \cos(n\tau - \phi))]$$

The parameter A controls the amplitude of this periodic function, n the periodicity, and ϕ shifts the entire curve along the rotation angle axis τ. Again the parameters for all combinations of four atoms have to be determined (e.g. C-C-C-C, C-O-C-C, H-C-C-N).

Non-bonding Energy:
The different implementation of force field differ mainly in the definition of this term. Mostly present are Van der Waals and electrostatic terms.

$$E = \underbrace{\sum_i \sum_j \frac{-A_{ij}}{r_{ij}^6} + \frac{-B_{ij}}{r_{ij}^{12}}}_{\text{Van der Waals}} + \underbrace{\sum_i \sum_j \frac{q_i q_j}{r_{ij}}}_{\text{Coulomb}}$$

The Van der Waals term accounts for the attraction and the Coulomb term for electrostatic interaction. Repulsion occurs, when the distance between two atoms becomes less than the sum of their radii. The shown approximation for the van der Waals energy is of the Lennard-Jones

potential type. It is used this way for instance in in the AMBER force field [65] as can be seen in equation 1. The last term accounting for H-bonds is modeled by a 6-12 potential as well.

$$
\begin{aligned}
E_{\text{total}} = & \sum_{\text{bonds}} K_r(r - r_{eq})^2 \\
& + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 \\
& + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\
& + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \\
& + \sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right]
\end{aligned}
\tag{1}
$$

2.3 Knowledge Based Potentials

In contrast to the analytic approach of mechanical force fields, knowledge based potentials describe the energy needed for a certain contact to occur by a likelihood. This likelihood of finding a particular contact is extracted from a database of known structures. Computer scientists would call this procedure *data-mining*. The increase of information is measured by the log-likelihood ratio of the Bayesian events [4]. This ratio is the relation of prior expected events and the observed occurrence. Therefore the log-likelihood is a kind of measure for the "surprise" provided by the database.

A physical interpretation of the probability function comes from statistical mechanics: Based on the assumption that the protein is in its energetic minimum, low energy elements must occur more frequently than others in $3d$-structures of globular proteins. This dependence of occurrence on energy resembles a Boltzmann statistic:

$$f_{occ.} \sim \exp -E/RT$$

Here T is the conformational temperature and R is the gas constant. This similarity reveals, that if in principle the frequency of occurrence can be estimated, it is possible to gain access to the putative energy of a certain fold $\psi(S)$. This interpretation of knowledge based potentials was introduced by Manfred Sippl and is the basis for most of the contemporary potentials of mean force.

Recently Dill and Thomas stated severe critic on this approach of statistical potentials [63]. They intended to test how "extracted" energies correspond with

"true" energies by mimicking the extraction process on ideal lattice models and comparing the observed with the accurate energy of **HP** interactions. Their major points of critic for this model are that proteins are not seen as chains (either as gas composition) and the temperature applied to the Boltzmann device is meaningless. Further they try to show that the energies for a certain fold depend solely on clustering of polarity. These findings were put into theoretic framework recently by Neumaier's *"Nonuniqueness Theorem"* [44]. It has been shown, that empirical potentials obtained by extraction of equilibrium geometries can *never* reveal *true* energies. In particular, empirical potentials derived solely from databases of equilibrium data will never be useful for dynamical studies.The relevance of these results will be discussed later on.

2.3.1 Statistical Thermodynamics of Proteins or the Inverse Boltzmann Law

The so called "*folding postulate*" states, that "*In equilibrium the native state of a protein-solvent system corresponds to the global minimum of free energy*". This was demonstrated in the pioneer study performed by Anfinsen [1] in 1973. He was able to show, that by reducing and re-oxidating disulfide bonds in ribonuclease no loss of function occurs, i.e. that folding is a reversible process.

The peptide chains will be presented by C^{α} atoms to make the model easier, by no loss of generality. According to Bolzmann's law the probability $f(x)$ of finding a physical system in a particular state x in equilibrium is give by:

$$f(x) = \frac{1}{Z} \exp\left[-\frac{E(x)}{kT}\right]$$

Where k is the Boltzmann's constant, T the absolute temperature in Kelvin (Reference temperature) and Z is the partition function defined as:

$$Z = \int \cdots \int \exp\left[-\frac{E(x)}{kT}\right] dx$$

For discrete systems the integral may be replaced by the sum.

$$Z = \sum_{x=1}^{n} \exp\left[-\frac{E(x)}{kT}\right]$$

If energies of all states x are known, the probability density could be computed. On the other hand it is possible to obtain the energy if the density of states can be measured [55].

$$E(x) = -kT \ln\left[f(x)\right] - kT \ln Z \tag{2}$$

From equation 2 it is possible to calculate the energy of a particular distribution but it is impossible to get the Boltzmann sum Z, so an additive constant remains unknown. If the probabilities of a distribution are extracted from a database, the potential of mean force of interaction can be obtained. If $E(x)$ denotes the reference state of the system (averaged energy), the net potential for a given interaction γ can be computed by:

$$\Delta E_\gamma(x) = E_\gamma(x) - E(x)$$

or:

$$\Delta E_\gamma(x) = -kT \ln \left[\frac{f_\gamma(x)}{f(x)} \right] - kT \ln \frac{Z_\gamma}{Z}$$

and since Z_γ and Z do not depend on the state x, it is legitime to assume $Z_\gamma \simeq Z$, and therefore $-kT \ln \frac{Z_\gamma}{Z} \sim 0$. T is tied to the temperature of the NMR or X-ray measurement of the data.

$$\Delta E_\gamma(x) = -kT \ln \left[\frac{f_\gamma(x)}{f(x)} \right]$$

Due to the restriction of a limited number of observations it must be distinguished between the probability densities $f(x)$ or $f_\gamma(x)$ and the information obtained from the database $g(x)$ respectively $g_\gamma(x)$. It is reasonable however to approximate the reference state probability $f(x)$ with $g(x)$ since the overall number of interactions in the database is big enough (magnitude of 10.000). On the other hand the number of observations can be low for particular contacts, especially when considering higher order interactions. Therefore database size is crucial for the approximation of $f_\gamma(x) \approx g_\gamma(x)$.

So without knowledge of any specific interaction we have to assume $f(x) \approx f_\gamma(x)$ and expect $\Delta E_\gamma(x) \equiv 0$. Each information quantum derived form the database increases $f_\gamma(x)$, and the net contribution is twofold: (1) The relative energy of all states $\Delta E_\gamma(x)$ is increased and (2) the energy of a particular state $\Delta E_\gamma(t)$ is lowered. This means that if $f_\gamma(x) < 1$ the contribution to the overall energy becomes negative. When parameters for all configurations γ are extracted, a summation over all contributions yields the energy of sequence S for structure (ψ):

$$E(S, \psi) = \sum_\gamma E_\gamma(x)$$

2.4 Various Approaches to Knowledge-Based Potentials

Over the past years many different approaches to potentials of mean force have been made. The various potential functions are distinct in the definition as well as

in the order of interaction. Therefore different "resolutions" are used to define the energy functions. The spectrum reaches from an atomic resolution mode (Sippl) to simplified **HP**-patterns (Crippen), and a lot in between.

Munson *et al.* [41] were able to show that increasing the order of interaction improves the statistical significance of the terms. Starting with a highly significant one body term, that counts for the exposures of the residue, continuing to a pair potential term, that contributes for amino acid preferences (e.g. hydrophobic-hydrophobic interactions) independent of the burial status, one can clearly identify that multi-body interactions participate to a major extent the overall potential function.

2.4.1 Atom-Atom Potentials

The reversible energy required to bring two particles close to each other at constant volume is given by the potential of mean force or Helmholtz free energy of the system. It is related to the radial distribution function $g(r)$ by:

$$w(r) = -kT \ln[g(r)]$$

and can give insights to protein folding and the role of specific interaction in native structures (e.g. H-bonds). The distribution function for arbitrary sets of atom-atom interactions occurring in proteins can either be obtained by diffraction experiments, or they are extracted from a database of structures. The two functions turn out to be equal, if the distance distributions are similar. The knowledge based distribution function is accessed by the determination of

$$\rho_{ab}(r) = \sum_{ab} \delta(r - r_{ij})$$

as the sum over all distinct pairs ab within the radius r in a protein library. The observed density is compared with a bulk of non interacting particles to finally obtain the distribution function:

$$g_{ab}(r) = \frac{\rho_{ab}(r)}{\rho}$$

The potentials using these distribution functions are perfectly suited for a detailed analysis of spatial distributions of atom contacts along a protein chain [59]. To make use of an atom-atom based potential, one has to know the Cartesian coordinates for *all* residues in a poly peptide chain. Therefore this approach is of no use to solve the inverse folding problem, as targeted by our group.

2.4.2 Sippl's PROSA II Potential

Sippl also implemented a pair potential in his software package PROSA II [55, 13, 56, 57]. The program was designed to determine the correctness of an experimentally derived structure under use of a quality factor *score*. The potential function used is a superposition of a pair-potential and a surface potential:

$$W(x,\psi) = \sum_{i<j} W_\gamma \left[x_i, x_j, |i-j|; \mathbf{d}^\gamma_{i,j}\right] + \sum_i V_\gamma \left[x_i; \chi(i)\right] \tag{3}$$

The first term W_γ stands for the pair contribution, V_γ is the surface part of the potential and both terms depend upon the backbone atom type γ (C^α or C^β). The pair-potential is calculated between amino acids x_i and x_j, located at position i and j of the sequence x. $\mathbf{d}^\gamma_{i,j}$ is the Euclidean distance of the contributing amino acids. Using a particular surface term is caused by the observation, that a solvent-protein interactions can be used to model amino acid energies more accuratly [8, 9, 10]. The parameter χ represents a quantitative measure for the extent of surface exposure of amino acid x. The potential function as described by the parameters $W_\gamma \left[x_i, x_j, |i-j|; \mathbf{d}^\gamma_{i,j}\right]$ and $V_\gamma \left[x_i; \chi(i)\right]$ are extracted from a representative pdb-subset, applying the Boltzmann principle, and distributed with the PROSA II-package.

Because PROSA only uses the C^β (or C^α) atoms of the backbone, and calculates the probability of finding two *residues* within a spatial distance, it is quiet well suited for inverse folding studies.

2.4.3 Lapedes' Neural Network NN Potential

Alan Lapedes *et al.* [26] developed a potential with multi-body interactions, parameterized in "local neighborhoods" for each residue.He generalized other threading approaches, and ended up in a statistical interpretation. To employ a neural net for finding a log-likelihood ratio containing higher order terms of interaction, it is necessary to find a suitable representation of the available structural information. To tackle this problem an internal coordinate system is defined, setting the C^α-atom to the center, and constructing two vectors pointing to the neighboring chain atoms: C and N. This plane has been shown to have an almost constant angle, and a third dimension is spanned by the cross product of $\overrightarrow{C^\alpha N} \times \overrightarrow{C^\alpha C}$. Further a binned sphere is constructed around the center (C^α-atom) of the coordinate system, representing a "neighborhood shell" of residues. To order this shell to spatial residues, the sphere is split into a predefined number of finite, binned sub-shells.

The chain neighbors, carrying information necessary for secondary structure, can be included as well. The M bins are filled with integers mimicking the 20 amino

acids, describing the surrounding of a particular C^{α} atom. The neural net is trained on the pdb-select database, and parameters as number of sub-bins, bin size, or bin resolution were varied. Approaches for C^{β} as a core atom showed better results in threading experiments.

2.4.4 Contact Potentials

Contact potentials can be understood as subgroup of knowledge based potential. This kind of mean energy function measures the overall energy of a system, as the sum of *nearest neighbor* contacts. The most prominent examples are:

Crippen's Simplified Potential

To obtain a simplified representation of heteropolymers Ken A. Dill introduced the concept of lattice polymers [14]. When used to model proteins, each amino acids occupies one positions on the grid of the lattice. Conformations of lattice polymers are represented by *self-avoiding walks*, short SAWs. Hence this method greatly reduces the conformational space of the optimization problem. On a lattice bond lengths are, of course, always constant, furthermore potentials for lattice proteins usually neglect bond angles and dihedrals. Instead they focus on non-bonding interactions of topological neighbors.

In Crippen's potential the energy for the pair interaction is written as:

$$E(s, \mathbf{x}) = \sum_{i,j} \Psi[s(i), s(j); |i-j|; d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j)]$$

The individual interaction terms Ψ depend on the type $s(i)$ and $s(j)$ of residues, on their separation $|i-j|$ along the chain and on the eucledian distance $d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j)$ of the lattice points. The potential function

$$\Psi[s(i), s(j); |i-j|; d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j)] = U[s(i), s(j); |i-j|]g(d_{\mathcal{L}}(\mathbf{x}_i, \mathbf{x}_j))$$

is normalized such that the contribution of the nearest neighbor reduces to $U[s(i), s(j); |i-j|]$.

Crippen extracted a contact matrix of the form:

$$U[s(i), s(j); |i-j|] = \begin{cases} -0.008 & \text{if } |i-j| = 3 \\ 0.004 & \text{if } |i-j| = 4 \\ 0.021 & \text{if } |i-j| = 5, 6, 7 \\ \begin{pmatrix} -0.012 & -0.074 & -0.054 & 0.123 \\ -0.074 & 0.123 & -0.317 & 0.156 \\ -0.054 & -0.317 & -0.263 & -0.010 \\ 0.123 & 0.156 & -0.010 & -0.004 \end{pmatrix} & \text{if } |i-j| \geq 8 \end{cases}$$

from a structural database where the matrix entries correspond to the four amino acids classes:

1 = {**G Y H S R N E**}
2 = {**A V**}
3 = {**L I C M F**}
4 = {**P W T K D Q**}

A further simplification of the potential can be obtained by restricting the amino acid alphabet to just two classes: **H** for hydrophobic amino acids and **P** for polar residues. For a review of **HP** based potentials see [15, 19]

Crippen recently used the described potential in kinetic simulations and calculations of denaturation curves [18]. These computer experiments showed, that folding kinetics largely depends on the coding scheme and that the results obtained by using the Crippen alphabet differs strongly from calculations for spin-glass encoded SAWs [23, 25].

Tropsha's Four-Point Potential

Avoiding the arbitrariness of a binned distance, A. Tropsha [68, 54, 69] introduced an approach from computational geometry to knowledge based potentials. He suggested to represent the protein structure as a set of points in $3d$, for simplification only C^{α} atoms were chosen as model for the backbone. This set of points is tessellated using the *Delauney triangulation*. The result of this geometric procedure is a partitioning of the space included by the set into irregular tetrahedra with the points as vertices. The quadruple of amino acids represented by these points are considered to be nearest neighbors. The beauty of this method is that it is parameter free, the list of tetrahedra is non-ambiguous.

If one counts the occurrence of all possible neighborhood combinations of the amino acids in a structural dataset, a log-likelihood function can be constructed

easily. This function can then be used to test if a given sequence yields favorable contacts when threaded to a certain structure — in one word *inverse folding*.

Since the implementation of a Tropsha-based potential is the core part of this work, it will be discussed in section 2.7 in depth.

2.4.5 Profiling Potentials

Eisenberg and coworkers decided to "translate" the $3d$-structures to a $1d$-string, using three parameters:

1. The total side-chain area being covered by any other protein atoms
2. The fraction of side-chain area being covered by polar atoms or water molecules
3. The local secondary structure

The environment strings were extracted from a database of known structures. The resulting environment classes discriminate buried and exposed residues, and further subdivisions yield 18 distinct classes for the 20 amino acids. The optimization problem was to find the most favorable alignment of a protein sequence to the environment string, whereby classical alignment techniques came to use. The resulting threading procedure has been successfully employed to identify sequence-structure pairs.

2.5 Delauney Tessellation

The common meaning of *"tessellation"* is to arrange squares in a mosaic pattern. The term derives from the Greek word *"tesseres"* which means four. Generally *tessellating* can be understood as arranging regular polyhedra congruently (all angles and sides are equal) in a plane with edges attached to each other. Only three regular polygons tessellate in the Euclidean plane: triangles, squares and hexagons (see figure 2). By extension, space or hyper space may also be tessellated.

The Delauney triangulation *tessellates* a set of points in $\mathbb{R}^3$ in the sense of filling space with tetrahedra. The Delauney triangulation is computed via its dual, the Voronoi diagram.

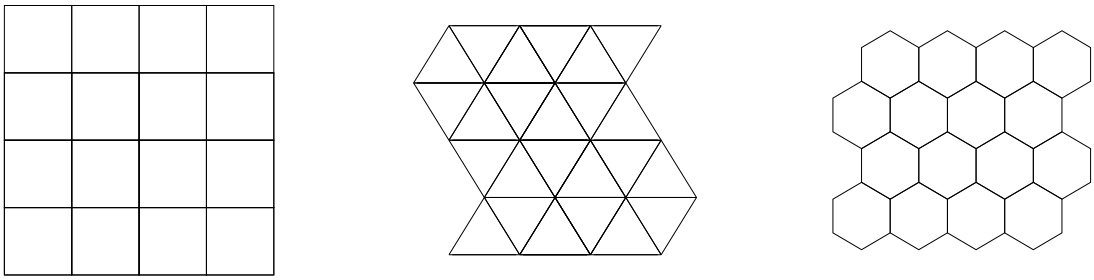


FIGURE 2: Tessellations in two dimensions.

2.6 Delauney Triangulation and Voronoi Diagrams

2.6.1 The Voronoi Diagram

Given a set S of n distinct points in $\mathbb{R}^d$, a Voronoi diagram is the partition of $\mathbb{R}^d$ into n polyhedral regions $\text{vo}(p), (p \in S)$. Each region $\text{vo}(p)$, called the Voronoi cell of p, is defined as the set of points in $\mathbb{R}^d$ which are closer to p than to any other points in S, or more precisely,

$$\text{vo}(p) = \{x \in \mathbb{R}^d | \text{dist}(x,p) \leq \text{dist}(x,q) \forall q \in (S-p)\}$$

where dist is the Euclidean distance function. The set of all Voronoi polyeders forms a cell complex. The vertices of this complex are called the *Voronoi vertices*, and the extreme rays (i.e. unbounded edges) are the *Voronoi rays*.

For each point $v \in \mathbb{R}^d$, the *nearest neighbor* set $\text{nb}(S,v)$ of v in S is the set of points $p \in S - v$ which are closest to v in Euclidean distance. In order to compute the Voronoi diagram, the following construction is very important. For each point p in S, consider the hyperplane tangent to the paraboloid in $\mathbb{R}^{d+1}$: $x_{d+1} = x_1^2 + \cdots + x_d^2$. This hyperplane is represented by $h(p)$:

$$\sum_{j=1}^{d} p_j^2 - \sum_{j=1}^{d} 2p_j x_j + x_{d+1} = 0$$

By replacing the equality with inequality $\geq$ above for each point p, we obtain the system of n inequalities, which we denote by $b - Ax \geq 0$. The polyhedron P in $\mathbb{R}^{d+1}$ of all solutions x to the system of inequalities is a lifting of the Voronoi diagram to one higher dimensional space. In other words, by projecting the polyhedron P onto the original $\mathbb{R}^d$ space, we obtain the Voronoi diagram in the sense that the projection of each facet of P associated with is exactly the voronoi cell $\text{vo}(p)$. The vertices and the extreme rays of P project exactly to the Voronoi vertices and the rays, respectively.

2.6.2 Delauney Triangulation

Let S be a set of n points in $\mathbb{R}^d$. The convex hull $\text{conv}(nb(S, v))$ of the nearest neighbor set of a Voronoi vertex v is called the Delauney cell of v. The Delauney complex (or triangulation) of S is a partition of the convex hull $\text{conv}(S)$ into the Delauney cells of Voronoi vertices.

The Delauney complex is not in general a triangulation but becomes a triangulation when the input points are non-degenerate, i.e. no $d+2$ points are cospherical or equivalently there is no point whose nearest neighbor set has more than $d+1$ elements. The Delauney complex is dual to the Voronoi diagram in the sense that there is a natural bijection between the two complexes which reverses the face inclusions.

There is a direct way to represent the Delaunay complex, just like the Voronoi diagram. In fact, it uses the same paraboloid in $\mathbb{R}^{d+1} : x_{d+1} = x_1^2 + \cdots + x_d^2$. Let $f(x) = x_1^2 + \cdots + x_d^2$, and let $\tilde{p} = (p; f(x)) \in \mathbb{R}^{d+1}$ for $p \in S$. Then the so-called lower hull of the lifted points represents the Delauney complex. More precisely, let

$$P = \text{conv}(\tilde{S}) + \text{noneg}(e^{d+1})$$

where e^{d+1} is the unit vector in $\mathbb{R}^{d+1}$ whose last component is 1. Thus P is the unbounded convex polyhedron consisting of $\text{conv}(\tilde{S})$ and any nonnegative shifts by the "upper" direction r. The nontrivial claim is that the the boundary complex of P projects to the Delauney complex: any facet of P which is not parallel to the vertical direction r is a Delauney cell once its last coordinate is ignored, and any Delauney cell is represented this way.

Considering a set of point in $\mathbb{R}^3$ the Delauney triangulation describes an algorithm to decompose the convex hull of these points into tetrahedra.

2.6.3 The `qhull` Algorithm

As previously described, the first step in generating the tessellation built from the irregular tetrahedron is finding the convex hull, which is the smallest convex set of points containing the entire set. The hull is represented by a set of facets and a set of adjacency lists giving the neighbors and vertices for each facet. In $\mathbb{R}^3$ facets are triangles and ridges are edges. The Delauney triangulation in $\mathbb{R}^d$ is calculated from a convex hull in $\mathbb{R}^{d+1}$ by lifting the points to a paraboloid by adding the sum of the squares of the coordinates and computing their convex hull, the set of ridges of the lower convex hull is the Delauney triangulation of the original set.

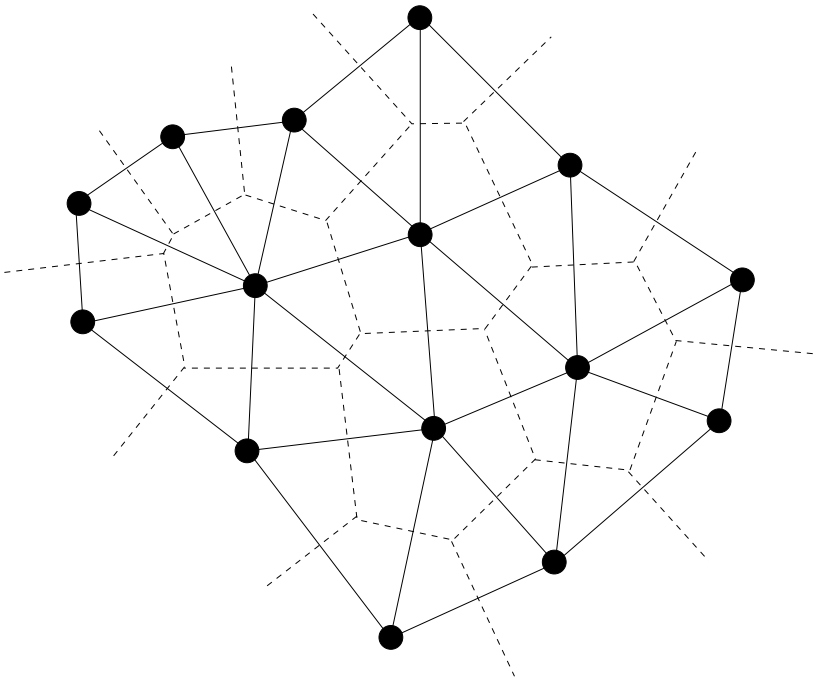


Figure 3: Voronoi diagram of a set of points in $2d$: The Delauney triangulation can easily be computed via its dual, the nearest neighbors of each Voronoi vertex are connected in the Delauney diagram. Voronoi cells are shown with dashed lines.

The `qhull` algorithm [11] is a variation of the randomized incremental algorithm, employing a constructed additional point at the hull to decide which facet belongs to it. The point is outside the facet if it is above the set and in the `qhull` variation of the original version, the point is not created randomly, but at the furthest distance from the outside set. This method is used in the program `qhull` which is publically available via the Internet [1)]. It has been shown empirically [11] that this algorithm is especially efficient and well suited for a $3d$ set of points.

This algorithm of triangulation can be applied to any set of points in space, always objectively describing neighborhood. Representing amino acids of a polypeptide chain by an atom (e.g. C^{α} or C^{β}) leads to a regular set of points in $3d$ space, that can be tessellated applying the rules described above. The Voronoi polyhedron is now the region around an atom, each side describes a contact to a neighbor. The underlying Delauney simplices are irregular tetrahedra with an amino acids at each corner. This diagram can be employed to describe contacts of amino acids objectively in $3d$ space.

[1)] URL: `http://www.geom.umn.edu/software/download/qhull.html`

2.7 Empirical Protein Potentials from Delauney Tessellation

2.7.1 Four Body Contact Potentials

Based on the fact that four *neighboring* points in space form an irregular tetrahedron, applying the Delauney tessellation to a set of points in $3d$ results in a contact potential. The likelihood of finding a distinct set of labeled points in this set can be expressed as:

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}} \tag{4}$$

where i, j, k, l are four amino acids, f_{ijkl} is the *observed* normalized frequency of occurrence of a given quadruple, and p_{ijkl} is the *a priori expected* frequency of occurrence of a given quadruple. So q_{ijkl} is a measurement of likelihood for finding four distinct amino acids in a simplex, namely a log-likelihood. The observed frequency f_{ijkl} is calculated by dividing the total number of occurrences of each quadruple by the number of all observed quadruples.

$$p_{ijkl} = C a_i a_j a_k a_l \tag{5}$$

where a_i, a_j, a_k, and a_l denote the individually observed frequency of occurrence of each amino acid residue. That is the total number of occurrence of a distinct amino acid type divided by the total number of residues in the dataset. C is the combination factor, acounting for the fact that replicated residue types are underestimated due to permutability. C is defined as:

$$C = \frac{4!}{\prod_i^n t_i!} \tag{6}$$

with n being the number of distinct residue types in a quadruple and t_i is the number of amino acids of type i.

Applying this procedure to a predefined set of experimentally derived protein structures leads to a potential of mean force. The calibration dataset has to be selected with care, since this selection determines the discriminative power of the force field. Parameters like the protein type (e.g. globular, membrane, soluble etc.), the type of backbone atom used for tessellation and any kind of selection of tetrahedra (i.e. filtering) have to be kept constant for the parameter set.

2.7.2 Energy and z-score

Statistical Analysis of the Delauney tessellation of a protein yields the q factors for the occurring quadruples as the likelihood of finding this particular contact

within the structure. Based on equation 4, it is possible to define the energy of a sequence x on a fold ψ as the sum over the log-likelihoods of all contacts that occur in ψ:

$$W(x, \psi) = \sum_{contacts} q_{contact} \tag{7}$$

where $q_{contact}$ is the statistic likelihood of a quadruple.

However since the determination of the ground state for each sequence would require to solve the folding problem, it is not possible to normalize the energy function. But defining a quantity called z-score as an energy separation between the native fold and the average of an ensemble of misfolds in the units of standard deviation of the ensemble, can be used for constructing an energy scale by which conformations between different sequences can be compared. Following Sippl [55, 13, 56, 57] we define

$$z(x, \psi) = \frac{W(x, \psi) - \overline{W}(x)}{\sigma_{W(x)}} \tag{8}$$

where $\overline{W}(x)$ is the average energy of sequence N in all conformations of a database and $\sigma_W(x)$ denotes the standard deviation.

The database has to be a source of alternative conformations for the sequence S with length n. If the database size x of possible structures is set to a fixed number, the number of possible decoys is a function of the sequence length. So for the limit $l \to N$ the database becomes insignificant. This problem has been circumvented by the construction of a "polyprotein" by linking all structures that are initially constructed for the measurement of the log-likelihood.

The sequence of the protein to be tested is slid along this aggregate of proteins from the N- to the C-terminus of the structural library amino acid by amino acid. For each aligned structure a z-score is calculated and counted as "misfold" to the ensemble, therefore it does not make too much sense to use a member of the dataset for testing the threading capabilities of the potential via the z-score . If $n \sim 40.000$ is the length of the poly-protein, $n - l$ missfolds can be constructed. Since $n \gg l$ this number of sequence-structure pairs is in the magnitude of the poly-protein length. This computational brute force attack is sufficient if it is not necessary to have gaps within the sequence-structure alignment. Otherwise more sophisticated techniques must be used.

An experimental test of the z-score [67] using thermodynamic data could demonstrate the definite significance of the scale. A z-score range from 15-30 for small native proteins could be observed. The magnitude of these scores shows the need to improve existing potentials. The scores derived form existing potentials are in range of 5-20, so they are too low.

2.8 Reduced Alphabet Potentials

Especially when asking for the origins of life it seems naturally to assume that prior stages of todays cells did not use the whole repertoire of 20 amino acids. Hypothesis about the origin of the genetic code postulate a *reduced* alphabet size as well since complexity always developed in steps.

On the other hand restrictions in alphabet size dramatically changes the energy landscapes, that give rise to the folding kinetics [66]. Considering the popular two-letter approximation:

H = {**A,C,I,L,M,F,W,Y,V**}
P = {**R,N,D,E,Q,G,H,K,P,S,T**}

it appeared that additional complexity is unavoidable. It has been shown by experiment that a protein domain (SH3) could still fulfill function when restricted to 95% **I**, **K**, **E**, **A** and **G**. This additional complexity must have been inevitably for fine tuning, otherwise it would be difficult to justify the use 20 letters in contemporary translation mechanisms.

An important feature for empirical potentials arising from reduced alphabets is the fact that the number of parameters is reduced dramatically, what leads to better statistics for observing particular contacts. The problem of known reduction schemes classifying the 20 amino acids is a lack of objectivity. The number of possible properties is enormous and largely "chemical" knowledge is applied, which mostly means any combination of parameters like hydrophobicity, acidity or charge. Therefore the classification depends more than less on the taste of the author.

Examples of coding are for instance the Crippen scheme (see page 13) or the one taken from Goldstein *et al.* [23], using 6 letters for coding (reffered as 6l):

Table 1: coding scheme for the 6 letter alphabet

c	**C**
f	**F Y W**
h	**H R K**
n	**N D Q E**
s	**S T P A G**
v	**M I L V**

The coding used by Crippen and this 6 letter alphabet shows no relation, both authors use different biophysical properties for grouping the amino acids.

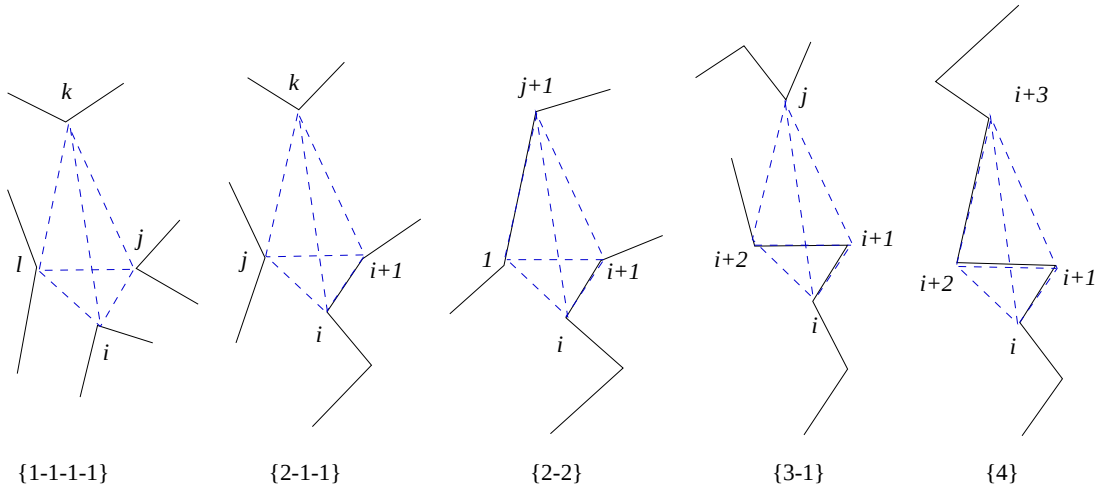


Figure 4: The five distinct Delauney classes as introduced to define the neighborhood in the amino acid chain for the simplices. The classes from left to right are:

class 0	all four residues are distant from each other	e.g. . 1-3-5-7
class 1	two residues are consecutive, the rest is distant	e.g. . 1-2-5-7
class 2	two pairs of neighboring residues	e.g. . 1-2-5-6
class 3	three amino acids are consecutive, one is distant	e.g. . 1-2-3-5
class 4	all residues are consecutive	e.g. . 1-2-3-4

For the analysis of sequence-structure correlation five additional classes were introduced, grouping the contacts according to their chain position of the participating residues (compare [68]). These *Delauney classes* pay respect e.g. to the steric hinderance of certain contacts of consecutive residues. Information about local secondary structural properties is obtained, therefore the protein does not appear as gas like agglomerate any more. This helps to improove the biophysical correctness of the model to a great extend. Figure 4 shows how the classification is constructed.

3 Methods

3.1 Computational Details — Overview

The practical part of this work involved the development of tools for calibrating a knowledge based potential, applying the tessellation as proposed by Alexander Tropsha. Since earlier studies of our group revealed a lack of consistence when ranking sequences optimized by the original Tropsha potential with the `PROSA` package extensions to the promising statistical geometric approach were intended to be made (for details see section 4.3.2).

Performing the tessellation for a given set of points in $3d$ space always generates per definition the convex hull. This of course leads to a very smooth surface for the protein structure in the model. To circumvent the artifact a particular filtering procedure is applied to the tessellation. Tropsha's implementation is lacking this extention.

In the original version of the potential C^{α} atoms were used for representing a residue in the poly peptide chain, in contrast `PROSA` uses C^{β}, arguing that these coordinates are more sensitive to the side chain orientation. Therefore it seemed natural to include this option in the calibration, interpolating a *virtual* C^{β} for the glycine residue, since this amino acid lacks that position.

Another term used by `PROSA` is paying respect to the difference between residues in the bulk and at the surface. This extension originates in the profiling approaches, introduced by Eisenberg *et al.* [8, 9, 10] An energy term for the surface can be combined via a scaling factor with the contact term.

The steps needed for the calibration of a tessellation potential and the required tools are listed below.

1. Determination of the database content from the list as described at [31] and downloaded pdb-files from EMBL [2)].
2. Generation of a proper dataset by pre-processing of the raw pdb-files.
3. Calibration of the desired potential (either contact or surface potential).
4. Post-processing of the parameter set.
5. Calculation of z-scores for the protein chains building the database.
6. Identify proteins chains not being globular, soluble structures and exclude them from the initially used pdb sub-set.
7. Recalculate potentials for the cleaned dataset.

[2)] URL: `ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select/`

3.2 Selection of a Representative Dataset

For the proper calibration of a knowledge based potential it is crucial to select a representative non-redundant dataset. The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules, serving the global scientific community [6]. The archive contains atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data. The entries are of a specific defined format [3] structured in a header section and a section containing the coordinates.

In 1998 the pdb database contained about 8.000 entries of atomic coordinates for proteins (according to the *pdb-newsletter*). This number however does not represent the number of different structures, which is by far smaller, because there is a lot of redundancy within the database (e.g. more than 70 structures of immunoglobulins can be found). Statistical analysis however require non-redundant data, so Hobohm et al. [49, 30] developed an algorithm to extract a subset with maximum coverage and minimum redundancy. The protein structures in the selected dataset had to fulfill the following requirements:

1. No pairs of proteins in the set have more than a prescribed level of sequence similarity.
2. The experimental quality meets certain criteria
3. The chains should not be shorter than a certain length

The question if two proteins are "close to each other" means that they are neighbors in sequence space. To generate a non-redundant subset one has to align each chain with all other members of the database. The arbitrary cutoff of sequence similarity is set to 25%. It has been shown by [49] that the exact value of this cutoff has only a weak influence, but for achieving a set of sequences that are widely spread in sequence space, the boundary has to be low. Another important point is that only high resolution structures are accepted as members of the selected set. This list is updated periodically and free accessible at EMBL [4].

3.3 Preprocessing of the Database

Prior to calibration of the potential the pdb-files were preprocessed with several tools, resulting in files only containing the amino acid backbone.

[3] URL: `http://pdb.pdb.bnl.gov/pdb-docs/Format.doc/Contents_Guide_21.html`
[4] URL: `ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select/`

First the files were split into separate files for each protein chain. The name of the chain (as given in the 21^{st} column of the `ATOM`-sector of the file, "_" if only one exists) was added to the file name. The chain was represented by the backbone atoms (C, O, N, C^{α}, C^{β}, if present) for a reduction of the file size. Since identical protein chains within one protein would distort the statistics these chains were ignored, if they are fully identical (sequence and length are equal). For theoretical models and models yielded by NMR-spectroscopy only the first model has been taken into account, discarding the rest. Furthermore nucleotides were omitted, and the chain is only processed if more than 30 atoms are present, since `qhull` needs more than 5 points to calculate a useful tessellation (either C^{α} or C^{β} atoms are processed). If an amino acid shows alternate locations (indicated by a letter in column 16 of the pdb entry) only version 'A' is used, all are others ignored. It is also an option to check for gaps within the chains, but the computational cost is quiet high, so this step is done while reading the files finally for tessellation.

Table 2: The June 98 – release of the *pdb-select* database shows the following characteristics:

Number of pdb files:	874
Total number of chains:	1663
Total no. of Amino acids:	357582
Average no. of chains/file:	1.9
Average no. of amino acids/file:	409
files rejected:	47
No. of C^{β} contacts:	613,170
No. of C^{β} contacts (unfiltered):	955,939
No. of C^{α} contacts:	557,252
No. of C^{α} contacts (unfiltered):	1,019,126
No. of contacts (C^{β} filtered):	8,847
No. of contacts (C^{α} filtered):	8,830

3.4 Tessellation and Counting of the Statistics

3.4.1 Construction of a Virtual C^{β} Atom

To enable calibration and calculation of scores for C^{β} atoms of a protein it is necessary to construct a virtual atom for glycine residues due to a lack of this position.

From figure 5 we see that the

$$\overrightarrow{X} = \frac{\overrightarrow{C^\alpha N}}{|\overrightarrow{C^\alpha N}|} + \frac{\overrightarrow{C^\alpha C}}{|\overrightarrow{C^\alpha C}|} \tag{9}$$

$$\overrightarrow{C^\beta} = \overrightarrow{C^\alpha} + h\frac{\overrightarrow{X}}{|\overrightarrow{X}|} + l\frac{\overrightarrow{C^\alpha N} \times \overrightarrow{C^\alpha C}}{|\overrightarrow{C^\alpha N} \times \overrightarrow{C^\alpha C}|} \tag{10}$$

$$\tag{11}$$

Equation 10 gives Cartesian coordinates for the "virtual" atom constructed. The values for h and l were measured at an alanine residues using *VMD* [35] and given in Ångstrøm.

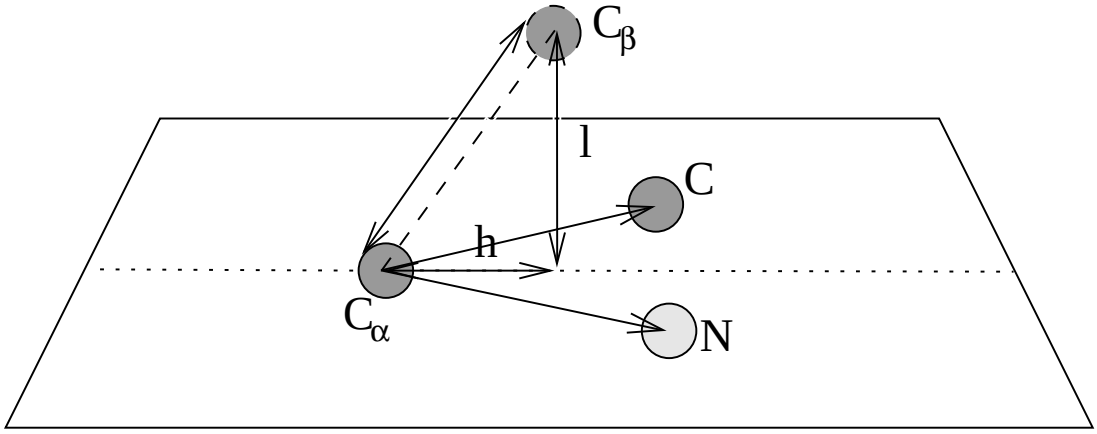


Figure 5: Construction of a virtual C^β by application of simple vector calculation and parameters obtained from alanine. The vector $\overrightarrow{X}$ in equation 9 is between $\overrightarrow{C^\alpha C}$ and $\overrightarrow{C^\alpha N}$ on the dashed line. The parameters $h = 0.88$ and $l = 1.24$ were measured.

3.4.2 Counting Statistics

If one wishes to recored all possible contacts between 20 different amino acids, $20^4 = 160,000$ terms would be generated. Assuming permutation symmetry of the vertices of each tetrahedron, the resulting categories reduce to a manageable amount of 8,855 different 4-tuples. As can be seen in table 2, not all possible contacts occurred in the database, since it is still very limited. The average number of contacts per quadruple is about 69.

Counting of the contacts is performed seperatly for all the various modes of tessellation (atom type C^α or C^β, filter on/off, alphabet size 20 or 6 letter) and the resulting parameters are written to an ASCII file. For calculation of the q_{ijkl} factors the relative frequencies of the amino acids have to be estimated as well.

Since not all 8,855 possible permutations of the amino acids tuples were observed, some log-likelihoods were calculated to be -Infinty. Those contacts were set to the minimum observed likelihood of that class in a post-processing procedure. A more rigorous procedure for this correction of sparse data still has to be developed.

3.4.3 Filtering the potential

Globular proteins are not necessarily convex. The tesselation hence may contain very flat tetrahedra or tetrahedra with unusual long edges. This problem arises from the construction of convex hull for the set, and is best illustrated by direct comparison of the surfaces obtained for filtered and unfiltered tessellations of a protein. As an example this was performed for the C^{α} atoms of Thioredoxin (pdb-id: `2trx`), the result is shown in picture 6. It can be seen, that all surface properties are lost without the filter. The filtering system introduced to the original potential identifies "bad" contacts by two parameters:

- The length of the edges of a tetrahedron must not be longer than 9.5 Å, what is a compromise due to loss of contacts, compare [41].
- Avoiding flat tetrahedra by setting a maximum circumsphere radius of 9.0 Å for the tetrahedra. A small radius suggests that the four residues are packed tighter.

Munson *et al.* [41] showed, that though applying this filter to the tessellation all relevant contacts still appear in the set. Unfortunately this true improvement introduces a parameter to the otherwise un-ambiguous approach.

3.4.4 Surface Generation

One of the most important innovations to the Tropsha potential is the use of a special term for surface contacts. This was mainly motivated by the observation, that with this potential optimized sequences showed unacceptable bad surface scores when cross-checked with the `PROSA` package.

The biophysical foundation is that proteins strongly interact with the surrounding solvent. Eisenberg and Bowie [8, 9, 10] demonstrated that solvent exposure can be used as a sensible parameter for the modeling of energetic features of protein-solvent systems. Reluctantly due to the mobility of solvent molecules only a small fraction of them can be monitored in X-ray experiments. Because of this lack of experimental data, an indirect approach has to be considered. Again employing the means of computational geometry can help to discriminate bulk and solvent phase.

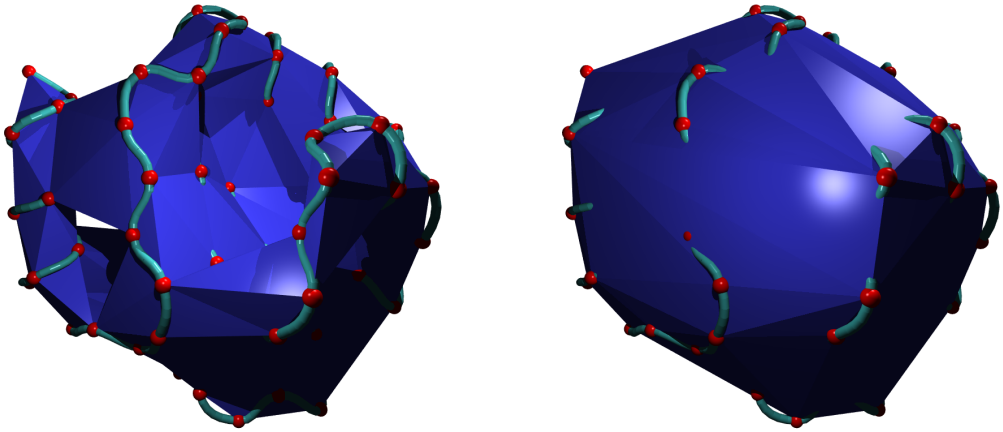


Figure 6: Comparing the filtered and unfiltered tessellation in the example of `2trxA`. The red balls represent C^{α} atoms, the green tube mimics the backbone of the protein. The right pictures shows the surface as generated by the Delauney Tessellation without any selection of tetrahedra. The left side was tessellated with the filter procedure as described above.

Contacts in $2d$ are unambiguously defined by the points of a triangle, therefore three amino acids form a contact that can be interpreted as surface. Trivially each tetrahedron generated by tessellation of the protein exposes 4 surfaces to the surrounding, either another tetrahedron of the package *or* the molecular surface as exposed to the solvent. Each triangle in the complete set only being member of one single tetrahedron therefore is considered as part of the surface.

Thus using a similar procedure as for obtaining the counting statistics of the tetrahedra can be applied to surface triangles, is, to those triangles that are contained in only one tetrahedron. member of the tetrahedra. The result again is a parameter set containing *log*-likelihoods for a specific triple of amino acids to appear in neighborhood and at the surface. The influence of filtering the surface is extremely high as expected , since a smoother surface provides fewer different contacts. The combined potential is defined as:

$$W^{comb} = W^{cont} + \gamma W^{surf} \tag{12}$$

with W^{cont} being the contact energy and W^{surf} representing surface energy for the appearance of three particular residues at the surface. The scaling factor γ weighs the the different influence of bulk and surface terms.

The surface energies as obtained by summation of all individual surface contact parameters is being combined with the contact energy via a scaling factor. This

factor pays respect to the order of interaction, and was *a priori* set to be 1. To test this assumption, the parameters are varied and plotted against the scores, the result is shown in figure 13 of section 4.

3.5 Iterating the Potential to Self-consistency

The nature of the *pdb-select* dataset makes it necessary to post-process the calibration dataset. Some of the z-scores for the native proteins in *pdb-select* were unacceptably bad when calculated with the `PROSA II` package (i.e. high in terms of the `PROSA` scale which is inverted). These files were removed from the calibration dataset, see table 3 for details.

Plotting z-scores from `PROSA` against the chain lengths (see figure 7) lead to the exclusion of some *pdb-select* entries because they obtained a score that did not show the observed dependency of the score and the chain length: usually longer chains get better scores, a linear curve-fit shows good accordance, those files being outliners were excluded as well and added to table 3.

Inspection of the excluded files showed in most cases that the basic assumption of globular, soluble proteins has been violated: in table 3 four main explanations are recorded to be the reason in most cases:

1. Membrane proteins are not soluble or globular in most cases: They have large parts that are surrounded by the hydrophobe part of the membrane.
2. Hydrophobic residues are also exposed on the surface if the sequence is not a complete protein.
3. The protein is part of a complex with other molecules(e.g. nucleic acids, or large prosthetic groups)
4. Another problem arises from the fact, that the *pdb-select* contains single chains from multi domain proteins. Viewing these chains isolated leads to the observation of denaturated proteins since it is easily possible to imagine that chains touch each other in hydrophobe regions. These regions would of course be water exposed after separation of the chains. This is especially important for surface potentials.
5. The structure is very long stretched and rod-like. Some of the entries contained coordinates for isolated α-helices, that did not obtain a good surface score.

Figure 8 shows a few examples for these artifacts.

The classification scheme used in table 3 is more than less arbitrary, but explains the outliner quite well. There is an overlap in the categories, the most obvious reason has been assumed to be relevant. The chosen cut-off value of scores ($z_{PROSA} > -3$) considered as "bad" was arbitrary, but under respect of a minimum in loss of files: see figure 9 for details.

Calculating the scores for the dataset using the tessellation potential yields a list similar to the one for calculated with `PROSA`. For convenience the cut-off was again set to be 3: Most files are the same, some were identified as "bad" that were considered ok by `PROSA`. This can be understood easily because the cut-off is taken arbitrary, and numerical values of the z-scores are of course not directly comparable between the two potentials.

This consistency is very promising, and a cross-check for the dataset within the tessellation potential showed a similar result. It could be observed, that especially the small proteins with bad scores were identified by both potentials in good correlation.

Proteins showing bad scores in one of the two potentials were excluded from the calibration dataset, since in all cases an obvious reason could be found why the score was bad. For future database procession an iterative consistency check as performed here will be done: folds showing a score smaller than a cut-off value are discarded (under consideration of the chain length dependency).

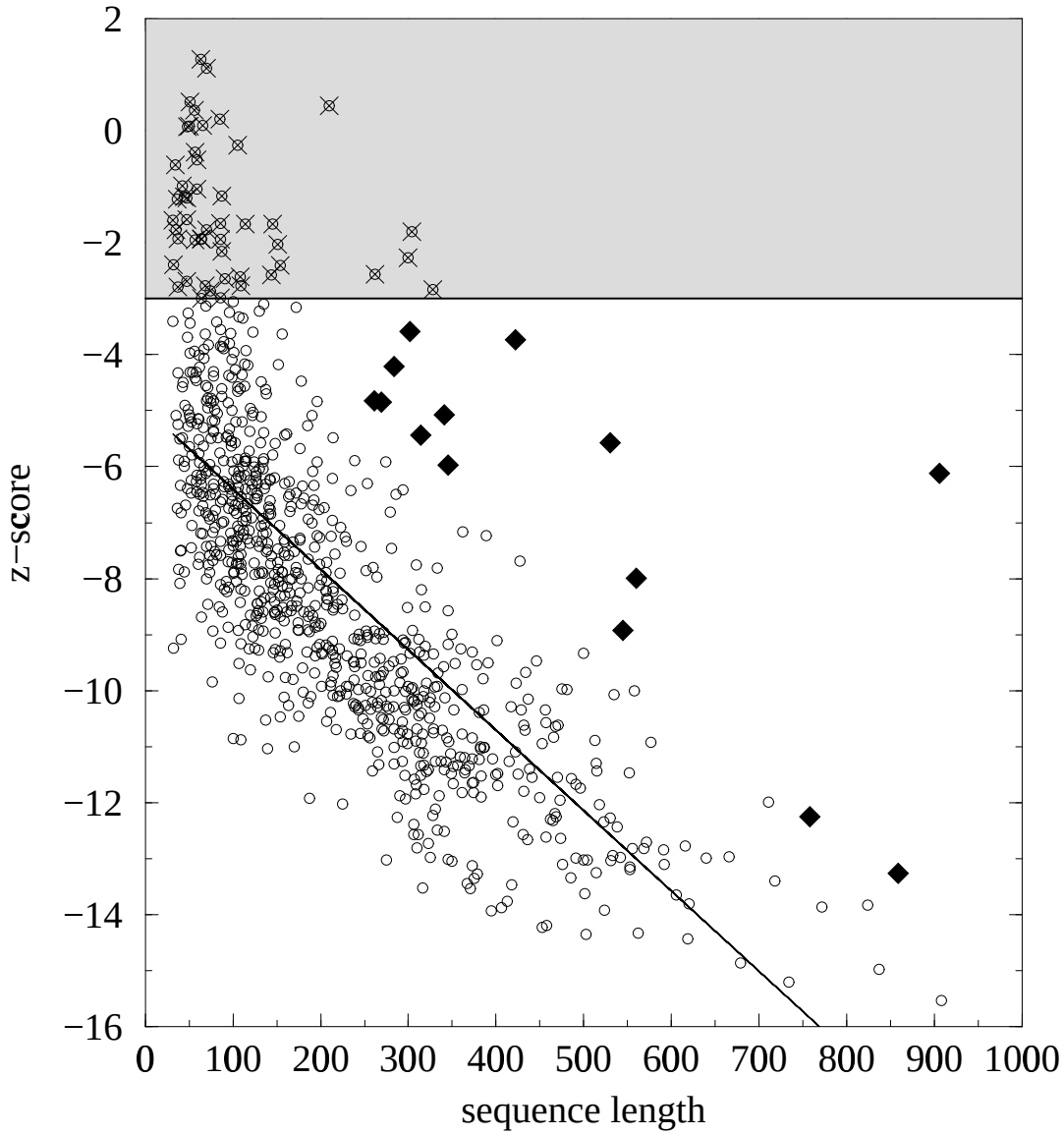
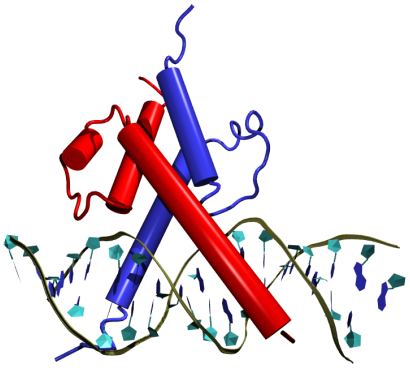
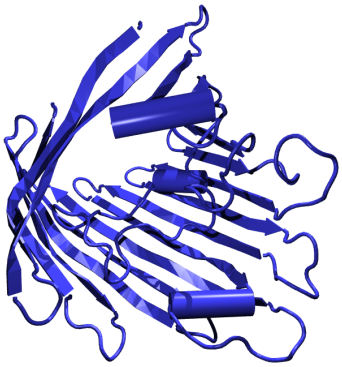


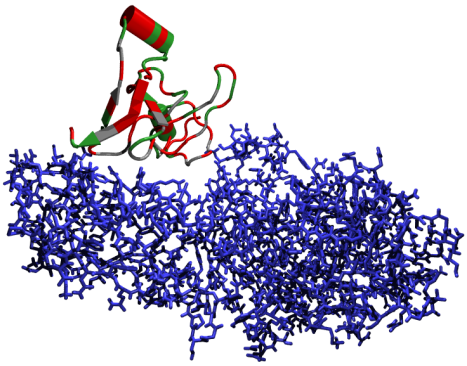
Figure 7: PROSA z-scores for proteins of the *pdb-select* database. Data points shown as diamonds represent proteins, that have a z-score too low for their length, since it has been observed, that usually longer chains yield better scores. Those points drawn as cross were excluded from the data set because they exceed the chosen threshold. The parameters from the regression are: $y = -0.01435x - 4.966$



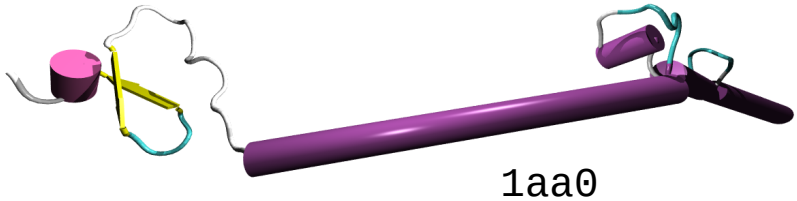
1a0a



2por



1lbp



1aa0

Figure 8: Examples of protein structures that should be excluded from the calibration data set:

- `1a0a`: This chain has a DNA ligand in its native structure
- `1aa0`: The structure is a long rod-like α-helix
- `1lpb`: The *pdb-select* contains isolated chains (here shown by its secondary structure), that expose hydrophobic residues if taken from the nativ counterpart (shown in blue licorice (hydrophobic regions are coded red in this plot)
- `2por`: Membran proteins are excluded *a priori*

Table 3: Proteins with PROSA z-scores higher than -3, the limit was chosen because a stricter judgment would result in a big loss of files (see figure 9 for details). Proteins with very low z-scores in the tessellation potential were excluded as well.

Protein (pdb-ID)	chain length	z-score (PROSA)	z-score (tessellation)	membrane protein	hydrophobe surface	non-protein ligand	multi chain complex or fragment	rod-like structure	Name of the protein
1a0a-a	62	-1.93	1.67			•	•		PHO4 BHLH Domain
1aa0	114	-1.67	2.15					•	Fibritin
1aaf	56	0.36	5.34					•	HIV-1 Nucleocapsid Protein
1aay-a	86	-1.94	3.46			•			Zinc Finger
1aie	32	-2.40	0.41					•	p53
1aij-s	300	-2.27	2.65		•				Photosynthetic reaction center
1aik-c	35	-1.77	0.37		•			•	HIV GP41 Core
1am7-a	151	-2.03	4.90				•		Lysozyme
1amm	545	-8.93	12.00		•				Gamma B-Crystallin
1aqz-b	144	-2.58	4.68						Restrictocin
1ar1-a	530	-5.57	6.47	•					Cytocrome C Oxidase
1as4-b	34	-0.61	12.00		•				Antichymotrypsin
1aty	47	-1.20	6.47		•		•		F1FO ATP Synthase
1auv-b	284	-4.21	11.36				•		Synapsin IA
1bbo	57	-1.95	2.90					•	Human MBP-1
1bcf-a	314	-5.44	11.36	•					Bacterioferritin
1bct	70	1.11	-0.72				•	•	Bacteriorhodopsin
1ben-b	31	-1.60	4.77				•		Insulin
1bhb	68	-2.77	2.00		•		•		Bacteriorhodopsin
1cfh	48	0.07	1.47					•	Coagulation Factor IX
1dhx	906	-6.12	6.98		•				Adenovirus Type 2
1fdm	51	0.51	0.74		•				FD Major Coat Protein
1fza-a	86	-1.66	1.96				•	•	Fibrinogen Fragment
1got-g	59	-0.52	0.59					•	GT-ALPHA/GI-Alpha
1hqi	91	-2.64	3.75					•	Phenol Hydroxylase Comp. P2
1hry-a	74	-2.86	1.65		•			•	Human SR
1htr-p	44	-1.18	1.94		•				Progastricsin
1hul-a	109	-2.77	3.02		•		•		Interleukin-5
1irk	304	-1.80	10.69	•					Insulin Receptor
1jsu-c	70	-1.77	1.91						Cyclin-Dependent Kinase-2
1kit	758	-12.25	16.12					•	Neuraminidase
1kzu-b	42	-0.99	1.37	•			•	•	Light Harvest. Compl.
1lgh-a	57	-0.38	1.08				•	•	Light Harvest. Compl. II
1lpb-a	86	-2.99	8.30		•			•	Lipase
1mnm-c	261	-4.82	4.66			•			MCM1 Transcriptional Regulaor
1msk	328	-2.84	9.19				•		Methionin Synthase
1myp-a	105	-0.26	2.47					•	Myeloperoxidase
1occ-c	262	-2.57	3.50				•	•	Cytochrome C Oxidase
1occ-d	114	-1.67	2.36				•		Cytochrome C Oxidase
1occ-g	85	0.21	1.31				•	•	Cytochrome C Oxidase
1occ-k	50	0.08	1.54				•	•	Cytochrome C Oxidase
1pdg-c	87	-2.16	7.15				•		Platelet-Derived Growth Factor BB
1ppt	37	-2.79	2.17					•	Avian Pancreatic Polypeptide
1qba	859	-13.26	4.65						Chitobiase
1tiv	87	-1.17	1.75			•			HIV-1 Transactivator

Table 3 continued.

Protein (pdb-ID)	chain length	z-score (PROSA)	z-score (tessellation)	membrane protein	hydrophobe surface	non-protein ligand	multi chain complex or fragment	rod-like structure	Name of the protein
1vba-4	63	1.27	1.51		•		•		Poliovirus
1vdf-a	47	-1.59	0.96					•	Cartilage Oligomeric Matrix Protein
1wdc-a	65	0.09	0.81					•	Scallop Myosin
1wht-b	154	-2.41	-1.23		•				Serine Carboxypeptidase II
1wpo-b	210	0.45	8.16		•				Human Cytomegalovirus Protease:
1ytf-c	47	-2.69	2.60			•			Yeast TATA-Box Binding Protein
1zto	37	-1.93	1.83	•					Potasium Channel Protein RCK4
1zwd	36	-1.23	0.32				•	•	Parathyroid Hormone Fragment 3-37
2drp-a	64	-1.94	2.53			•			Tramtrack Protein
2mev-1	269	-4.85	6.67				•		Mengo Encephalomyocarditis Virus Coat
2mev-4	59	-1.04	1.39				•	•	Mengo Encephalomyocarditis Virus Coat
2mpr-a	422	-3.73	6.96	•					Maltoporin
2omf	341	-5.07	6.64	•					OMPF Porin
2por	302	-3.59	4.07	•					Porin
2spc-a	108	-2.61	1.17	•					Spectrin
3bcl	345	-5.97			•				BacterioChlorophyll-A Protein
4dpv-z	560	-7.99	9.03	•					Canine Parvovirus Strain D Viral Prot. 2

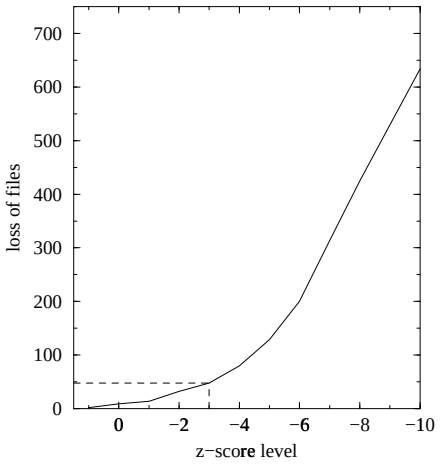


Figure 9: Distribution of the z-scores: The line shows the number of files excluded from the *pdb-select* database as a function of the treshold of the z-score (Data refer to the June 1998 release of the database)

3.6 Inverse Folding Using Knowledge-Based Potential

As stated initially, the hyper-astronomic size of sequences map into distinct areas of stable folds. The number of possible sequences exceeds the number of structures by far [3, 2]. Proteins frequently adopt similar $3d$-folds even if they are completely unrelated on the sequence level [46]. We define the neutral set of a native structure ψ to be the set of all sequences that fold into ψ according to the z-score criterion:

$$S(\psi) = \{x \in \mathcal{Q}_{20}^n \mid z(x, \psi) \geq z^*\} \tag{13}$$

where z^* means the z-score threshold level that must be reached by a sequence to be considered native-like. Inverse folding aims to identify sequences that fit into a distinct conformation, using an energy parameter as a guide (i.e. z-score). The solution of the inverse folding problem is by far more feasible than folding a sequence without a known structure, since shape space dimensions are immense, and optimizing a structure exceeds the computational possibilities by far and often structures derived from energy minimization, Monte Carlo and MD studies violate basic steric constrains. The optimization problem is very easy from the computational point of view:

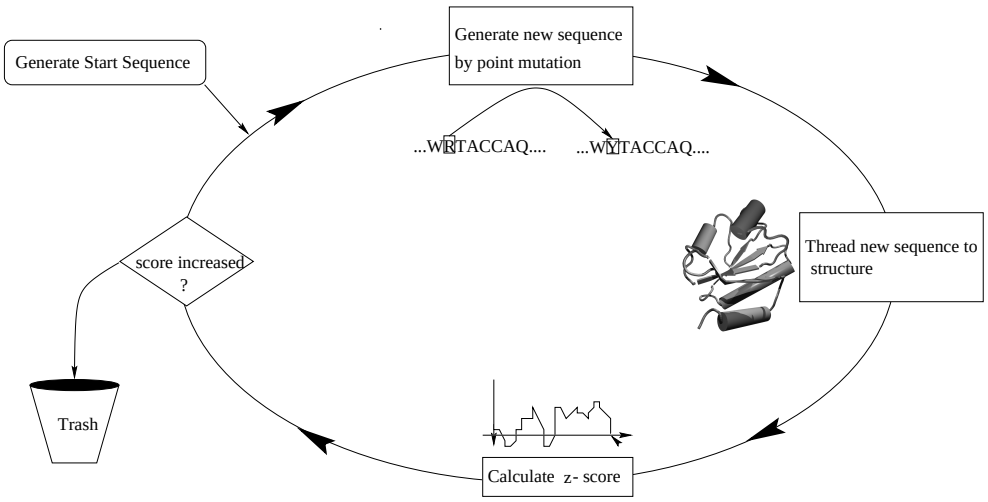


Figure 10: Schematic drawing of inverse folding using an adaptive walk to find the optimum

An *adaptive walk*, the simplest heuristic optimization algorithm, is sufficient. That means one repeatedly tries random point-mutations on the test-sequence, that are only accepted if they lead to an increase of the z-score. This would normally end up in a local minimum, but in practice a target z-score z^* is to be reached, in most cases that of the wild type sequence. For consistency sequences optimized with one type of empirical potential should obtain a good score with other methods of mean force as well. This has been shown to be true for the `PROSA` and `NN` potentials. For the tessellation potentials this test will be reported in the following

4 Results

4.1 Validation of the Potential

4.1.1 Re-Evolution with New Database

First a comparison of z-score calculated with the original data was done (table 4), showing that basically in most of the cases for the five proteins (which were arbitrarily selected) we obtain a better score within the new potential. This can be understood by the consideration that A.Tropsha used about 100 protein chains for the calibration of his set, while in our case about 700 chains were tessellated for the calibration. In general, if more information is provided as input for the q-factors, the energy contributions increas. All scores were calculated for C^{α} atoms and the same PolyProtein (`poly10k.pdb`). To apply the filter to tessellations using the original parameter set is incorrect because the the filter procedure was not used for the calibration by Tropsha.

Table 4: Comparison of the contact potential based z-score for new derived C^{α} tessellation potential data (E_{tess}, z_{tess}) with parameters obtained from original parameters as provided by A. Tropsha [54, 68] ($E_{Tropsha}$, $z_{Tropsha}$) for the six and 20 letter alphabets. z-scores printed *slanted* don't show the enhancement

Name	Alphabet	filter	$E_{Tropsha}$	$z_{Tropsha}$	E_{tess}	z_{tess}
`1bpi` (58)	20l	on	6.514	4.627	14.780	6.315
		off	3.125	3.616	4.709	*3.360*
	6l	on	7.805	4.618	17.845	6.766
		off	5.845	2.943	8.149	3.166
`2trxA` (108)	20l	on	9.823	4.709	16.586	5.623
		off	29.944	7.578	27.137	7.850
	6l	on	13.370	5.322	21.356	5.675
		off	38.622	8.460	6.097	*8.399*
`1bnr` (110)	20l	on	2.657	3.791	4.148	4.382
		off	0.609	3.860	2.044	4.026
	6l	on	4.836	4.303	4.019	*3.263*
		off	2.435	3.480	0.502	*3.168*
`1hab` (141)	20l	on	1.221	2.814	10.589	3.066
		off	14.592	5.242	9.524	*5.061*
	6l	on	1.499	2.343	12.461	3.962
		off	17.620	5.907	15.115	*5.786*
`1hjt` (153)	20l	on	4.861	2.774	15.178	5.097
		off	38.733	6.965	37.426	7.357
	6l	on	4.639	2.668	9.456	3.603
		off	42.733	8.646	39.595	8.665

As observed for the original parameters, reduced alphabets lead to better z-scores because of the improved statistics, but the effect of improvements are not so high as in the 20 letter case. Remarkably there was only one example where a filtered potential using the new parameters did not show an enhancement (`1bnr`, 6l). Figure 11 shows schematic drawings of the protein structures.

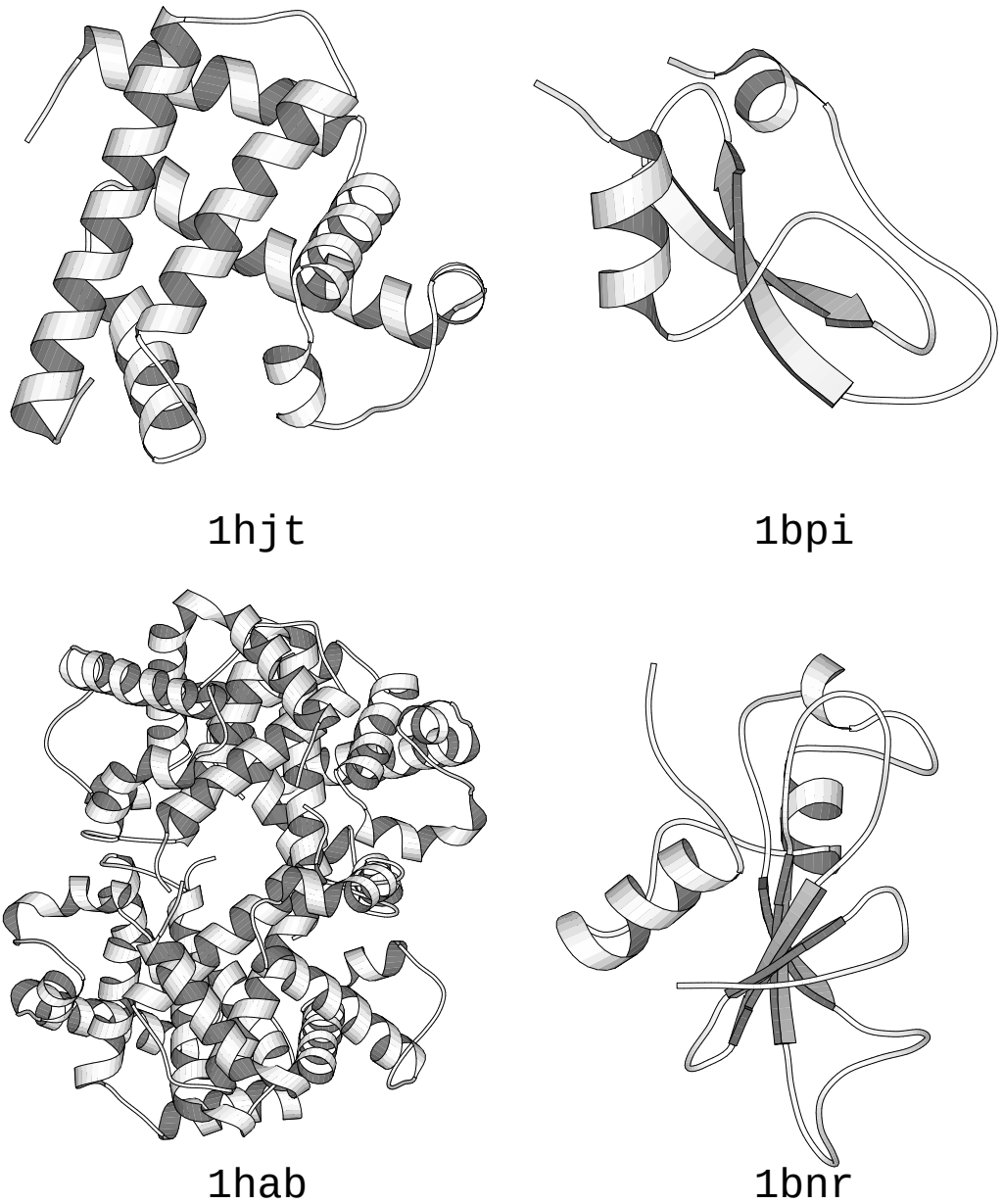


Figure 11: Pictures of the sample proteins used for the comparison of C^{α} and C^{β} potentials. The proteins are (starting from left top): `1hjt`: sperm whale myoglobine, `1bpi`: bovine pancreatic trypsin inhibitor, `1hab` haemoglobin, `1bnr` barnase).

4.1.2 Enhancements by extensions

The first innovation is the implementation of C^β atoms as representatives for the amino acid chain. The enhancement achieved is reasonable due to the fact that C^β is oriented sterical towards the side chains of the backbone. Table 5 shows the direct increase of the z-score if switched from a C^α to a C^β calibrated potential, regardless of the alphabet used. The same is true for the surface score. In all the cases the `pII3.0.short.pdb` PolyProtein was used, all calculations were performed using the filter.

Table 5: Comparison of C^α and C^β based potentials used for calculating z-score of native sequences

ID (length)	alphabet	atom	E_{cont}	z_{cont}	E_{surf}	z_{surf}	E_{comb}	z_{comb}
1bpi (58)	20l	C^α	14.780	6.315	-2.454	1.606	12.326	5.510
		C^β	19.161	6.758	-1.110	2.172	18.051	6.537
	6l	C^α	17.845	6.766	-0.558	1.792	17.287	5.837
		C^β	20.729	6.637	0.722	2.538	21.451	6.263
2trxA (108)	20l	C^α	16.586	5.623	6.216	5.909	22.802	7.224
		C^β	37.350	8.311	1.707	4.892	39.056	9.178
	6l	C^α	21.356	5.675	10.352	6.789	31.708	6.998
		C^β	46.400	8.544	6.750	6.477	53.150	9.322
1bnr (110)	20l	C^α	4.148	4.382	-0.323	2.588	3.826	4.861
		C^β	11.307	5.432	2.423	2.461	13.730	5.779
	6l	C^α	4.019	3.263	-1.007	1.474	3.012	3.036
		C^β	8.436	3.646	4.232	3.318	12.668	4.119
1hab (141)	20l	C^α	10.589	3.066	2.979	5.139	13.567	4.811
		C^β	22.972	5.306	1.216	5.679	24.188	6.661
	6l	C^α	12.461	3.962	9.440	6.976	21.901	5.769
		C^β	28.708	5.640	8.420	7.001	37.128	6.935
1hjt (153)	20l	C^α	15.178	5.097	13.374	6.541	28.553	6.990
		C^β	35.230	7.296	16.614	7.151	51.844	8.675
	6l	C^α	9.456	3.603	18.826	8.184	28.282	5.569
		C^β	33.480	5.637	22.059	9.000	55.539	7.138

In table 6 an exhaustive variation of all possible options has been performed. First C^α scores were calculated for the original potential. The parameters were originally derived *not* using a filter procedure by Tropsha *et al.* [54, 68, 69]. If filtering is applied to the tessellation while z-score evaluation the score decreases.

The newly calibrated potentials were employed for the calculations in the second table below. Generally C^β calculations provide better scores as C^α, but for C^α the

combined score is lowered by the surface score. The best z-score reached is in the range where scores of native proteins should be, in accordance with experiment. In all C^{β} calculations the use of the combined potential improoves the z-score .

TABLE 6: influence of all extension terms on the z-score of `2trxA`: All scores were calculated using the `poly10k.pdb` - PolyProtein. The first table shows scores for the original potential with variation of filter and alphabet size. In the second table the parameters used were generated under identical conditions as the z-score calculation.

Atom	filter	Alphabet	E	z_cont
CA	0	20l	29.944	7.730
	1		9.823	4.694
	0	6l	38.622	8.639
	1		13.370	5.390

Atom	filter	Alphabet	E_{cont}	zs_{cont}	E_{surf}	z_{surf}	E_{comb}	z_{comb}
CA	0	20l	6.598	8.616	18.472	4.136	55.070	8.185
	1		16.868	5.977	6.231	5.329	23.099	7.125
CA	0	6l	36.598	8.616	18.472	4.136	55.070	8.185
	1		21.639	6.076	10.372	8.193	32.010	7.636
CB	0	20l	28.267	8.237	30.755	5.769	59.023	9.206
	1		38.288	8.228	1.770	4.989	40.058	9.143
CB	0	6l	39.244	8.831	23.332	4.877	62.576	9.063
	1		46.726	8.488	6.683	6.479	53.409	9.280

4.1.3 Sequences Identify Their Structures

One important question concerning the potentials quality is if a natural sequence recognizes its native structure, and discriminates a different fold [29]. This feature can be tested by assigning pairs of proteins with the same length the sequences of each other. Table 8 shows the results for the corresponding calculations. The protein examples were essentially taken from [41].

The calculations were all performed using the 20-letter alphabet and using a potential calibrated for the *pdb-select* database released in July 1998. The potential as well as the z-score were generated applying the filter to the tessellation for the C^{α} atoms of the proteins. It can be seen, that in the given range of sequence lengths between 36 and 293 amino acids there is a severe discrimination between the corresponding pairs. The combined surface and contact potential show almost always an improvement of the distinction.

Table 7 shows an extended approach to thread natural sequences of length 108 as well as random sequences to the `2trx` structure of Thioredoxin. In all observed

cases the scores for any other than wild type sequence was out of the range of being considered wild-type like. The maximum difference to the wild-type z-score was about 10.

The comparison with PROSA score differences shows, that there is still room for improvement at the tessellation potential. The differences between native and non-native score are generally more distinct if PROSA was used for threading.

TABLE 7: Threading sequences of length 108 through 3d-structure of 2trxA

Name	E_{cont}	z_{cont}	E_{surf}	z_{surf}	E_{comb}	z_{comb}
2trxA	16.432	5.644	6.260	5.906	22.693	7.213
1cdp	-12.652	-0.752	-4.688	1.316	-17.340	-0.081
1cew	-9.128	-0.129	-8.133	-0.925	-17.261	-0.487
1chj	-19.448	-0.352	-3.254	0.825	-22.702	-0.007
1cih	-14.071	0.284	-6.330	-0.253	-20.401	0.148
1cri	-16.575	-0.213	-4.723	0.137	-21.298	-0.132
1rro	-25.261	-1.277	-5.619	-0.080	-30.880	-1.184
rand1	-20.243	-2.648	-5.118	-0.916	-25.361	-2.611
rand2	-16.707	-1.228	-3.202	-0.762	-19.909	-1.343
rand3	-11.772	-1.309	-8.898	-2.126	-20.670	-2.000
rand4	-18.632	-0.918	-5.398	0.501	-24.030	-0.619
rand5	-11.969	-0.298	-11.067	-2.038	-23.036	-1.107

TABLE 8: For pairs of example proteins with the same sequence length z-score were calculated for the native sequence of the structure and for the sequence belonging to the different protein of equal length. A comparison of z-score shows, that each sequence is able to identify its native structure and, combined potentials increase the discriminating power. The last columns show PROSA scores, to facilitate comparison the negative score is shown here.

Sequence		z_{comb}		$z_{contact}$		PROSA $-z_{comb}$	
PDB id.	l_{seq}	A	B	A	B	A	B
A: 1cbh	36	7.81	-0.98	7.78	-0.44	4.76	1.12
B: 1ppt		-2.68	1.67	-1.81	1.03	-0.13	2.79
A: 1fdx	54	5.65	-0.38	4.76	-0.00	7.46	1.29
B: 5rxn		-0.07	2.53	-0.23	2.92	1.38	5.87
A: 1ubq	76	8.81	-3.47	8.73	-2.38	9.25	-0.30
B: 4icb		-1.25	6.32	-1.29	5.26	1.05	10.06
A: 2trx	108	9.28	-1.36	8.48	-1.92	7.04	0.61
B: 1rro		-2.43	6.08	-2.20	4.87	-0.13	10.88
A: 1rhd	293	3.67	-0.59	2.88	-0.34	10.29	-0.46
B: 2cyp		-0.51	2.23	-0.49	7.26	0.85	10.21

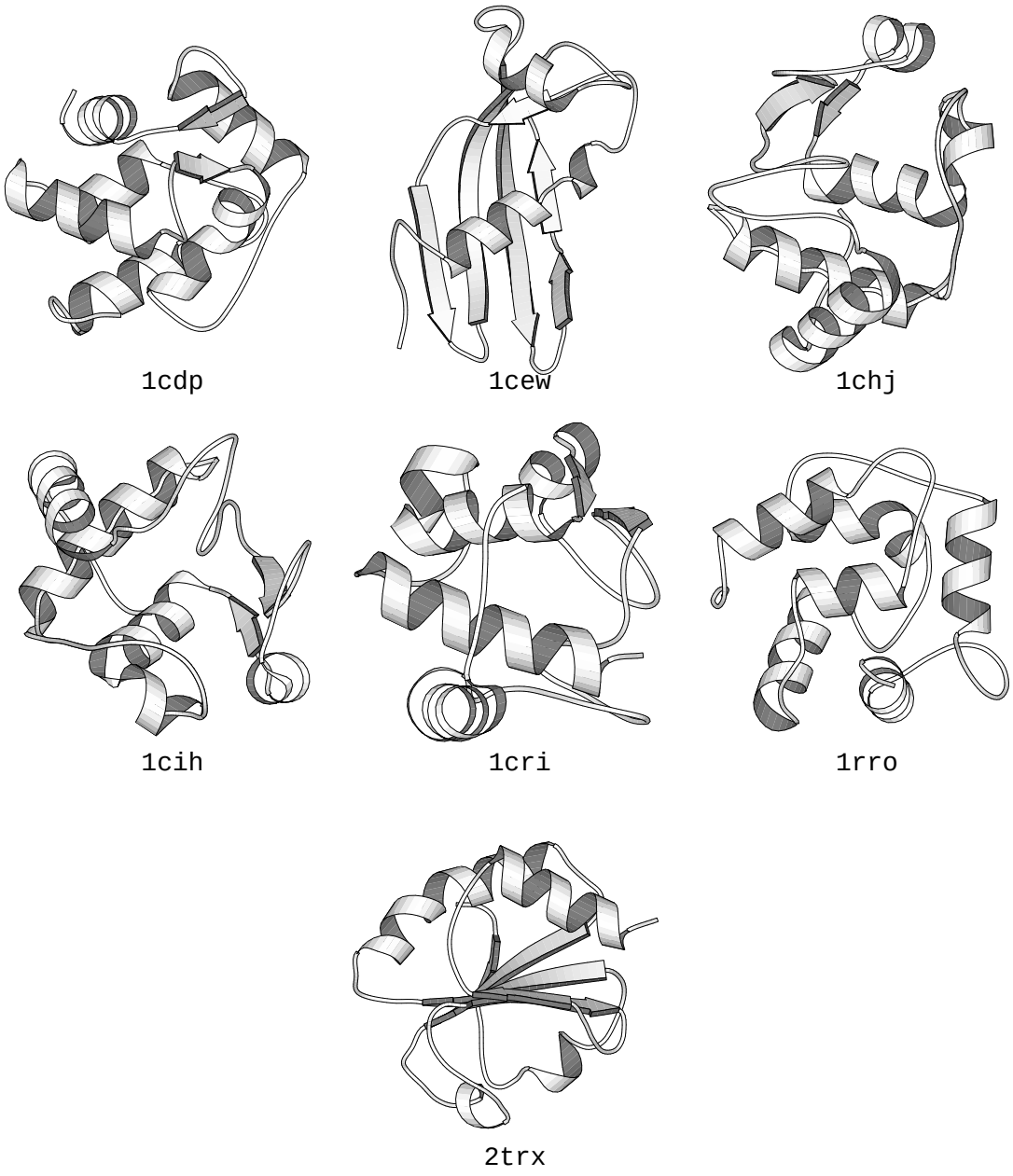


Figure 12: Proteins with length 108: threading the sequence of these structures into the `2trxA` structures shows that the potential is able to recognize the correct pair. The proteins are: (pdb-id's starting left top) `1cdp`: Parvalbumin B, `1cew`: cystatin, `1chj`: cytochrome C, `1cih`: cytochrome C, `1cri`: cytochrome C, `1rro`: rat oncomodulin, `2trxA`: thioredoxin)

4.1.4 Influence of the combining factor

Furthermore the weighting factor for the surface potential term has been tested. The influence shows nothing unexpected: increasing the factor emphasis the surface term inadequately until it is predominant. It is not appreciated to overestimate the surface contributions, therefore the combining factor in further calculations is set to 1. A plot showing the scores as a function of the factor for four different proteins is shown in figure 13.

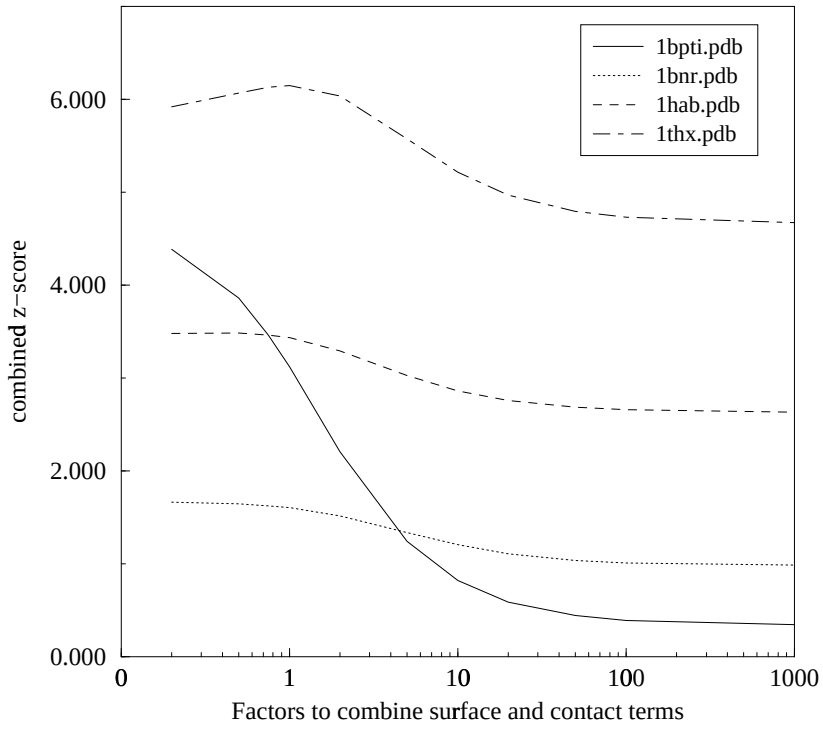


Figure 13: The variation of the combining factor shows nothing unexpected. First the contact term is predominating, with increasing weight the surface term supersedes.

4.2 Visualization of a four point potential

The result of the calibration process is a set of log-likelihoods for the distinct contacts. This (long) list of parameters contains all the information of the extracted database. An analysis of this data can show that known biophysical properties of proteins are indeed content of the empirical potential. For instance hydrophobe interactions of certain amino acids have to yield a higher liklihood for finding them in neighborhood.

The distribution of the q_{ijkl} values is shown as histogram for the 20l alphabet in figure 14. The plot is not scaled, therefore each peak represents the absolu-

te frequency of occurrence. The hull over all peaks is approximately Gaussian, with the maximum at 0 as expected. There are few outliers, indicating a good correspondence with the model.

Figure 15 shows the distribution of likelihoods for the six letter encoded potential.Each Delauney class has its own plot. Classes 0 and 1 show a wide standard deviation, in correspondence for contacts in distant chain regions. Only few observations could be made in class 4, what is reasonable since a contact of four serial amino acids is unfavorable due to steric reasons.

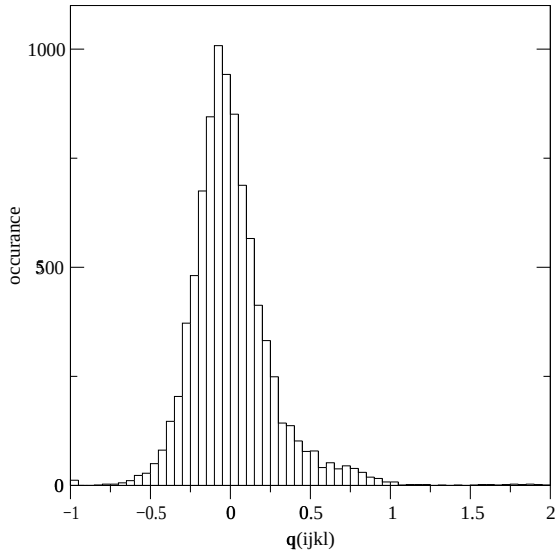


FIGURE 14: Distribution of q_{ijkl} values: the z-score for the **CCCC** - contact(3.24) is omitted since it is out of scale.

Long lists of parameters are ugly to look at and a lot of imagination is needed to extract relevant information. Because there is no obvious way to plot $4d$ data in a plane, I followed Munson *et al.* [41] and represented energies as graphical arrays. The plot is organized in a square, each side displaying an amino acid. this gives the first two of the four contacts. This huge square is further sub-divided into alphabet-size *times* alphabet-size sub-squares, each standing for a particular interaction of two residues, and partitioning the area into the contacts of four amino acids. The energy for the contact is color coded and normalized to be in a range between -1 and 1. For the representation all possible permutations of arranging the alphabet within groups of four are displayed. There was no observeable difference in the all-over pattern of the plot for C^{α} and C^{β} potentials.

The potential for the 20 letter alphabet is shown in figure 16, the color code in figure 17. The alphabet was arranged to have the hydrophobe residues at the beginning, therefor the down left corner shows the broad intensities of apolar clustering. The most intense contact to be found is the **CCCC**-quadruple, what

can be easily interpreted by disulfide bonds in the proteins. Clearly tetrahedra with both hydrophobe and hydrophil residues are hardly found.

The 6l alphabet is shown for all the Delauney classes in figure 19 for the newly calibrated potential and in figure 18 for the original Tropsha data. The patterns are similar in general: Classes 0 and 1 show a pattern similar to the 20l representation. The classes 3 and 4 representing the other end of the scale are distinctly different: e.g. v-clusters are highly favorable in class 0, whereas in class 4 they are unlikely to occur. Biophysically very clear seems to be that a **CCXX** contact ("**X**" stands for anf other amino acid here) in consecutive residues are hardly found. A comparison with the 6l5t potential data plot from the original parameters (figure 18) shows mainly that the number of all contacts increased (the overall plot is more intense in color) class 0 contacts for **CCXX** were additionally found in the new dataset.

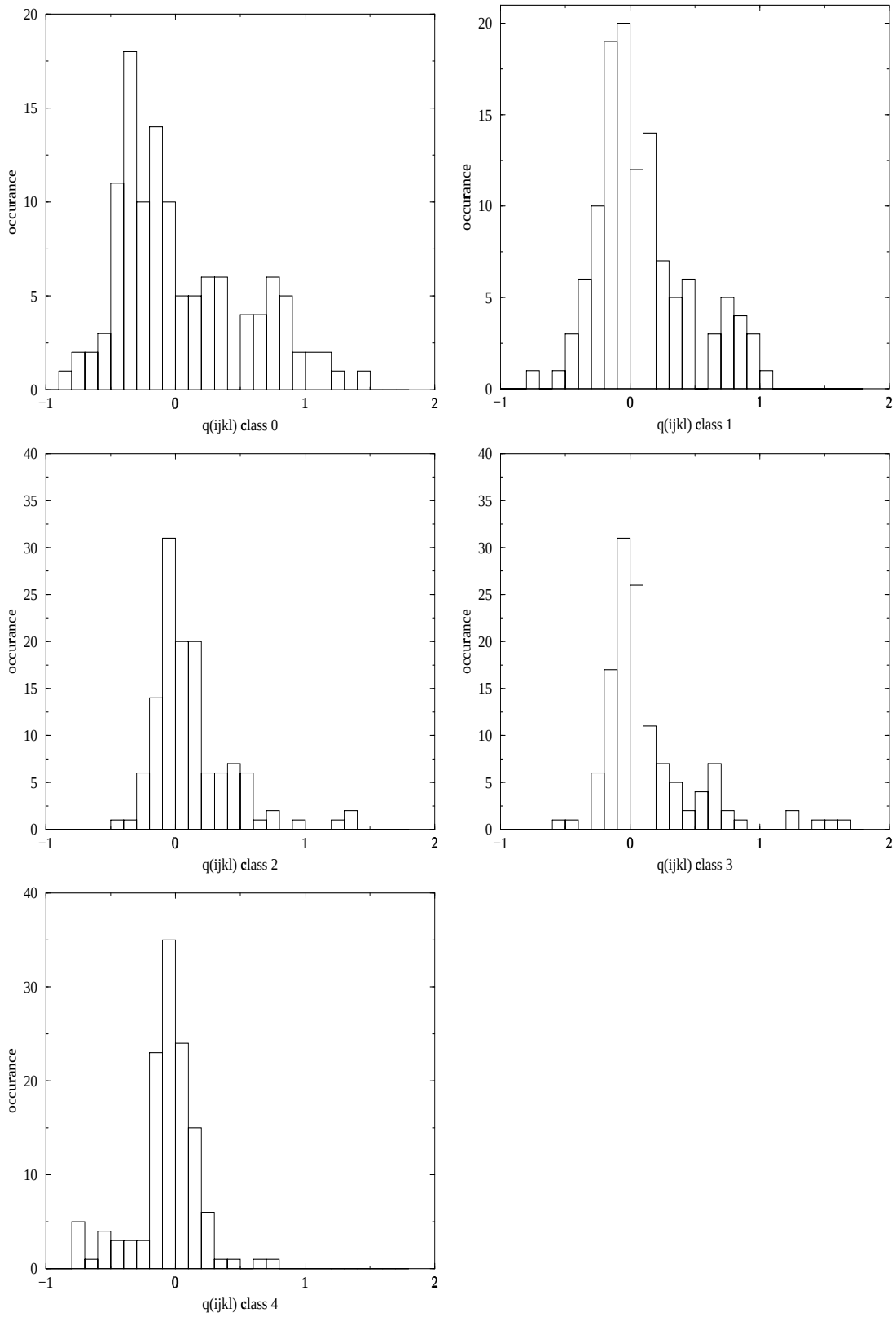


FIGURE 15: Distribution of q_{ijkl} values for the five Delauney classes of the 6l alphabet: class 0 means all residues are distant, class 4 means residues consecutive

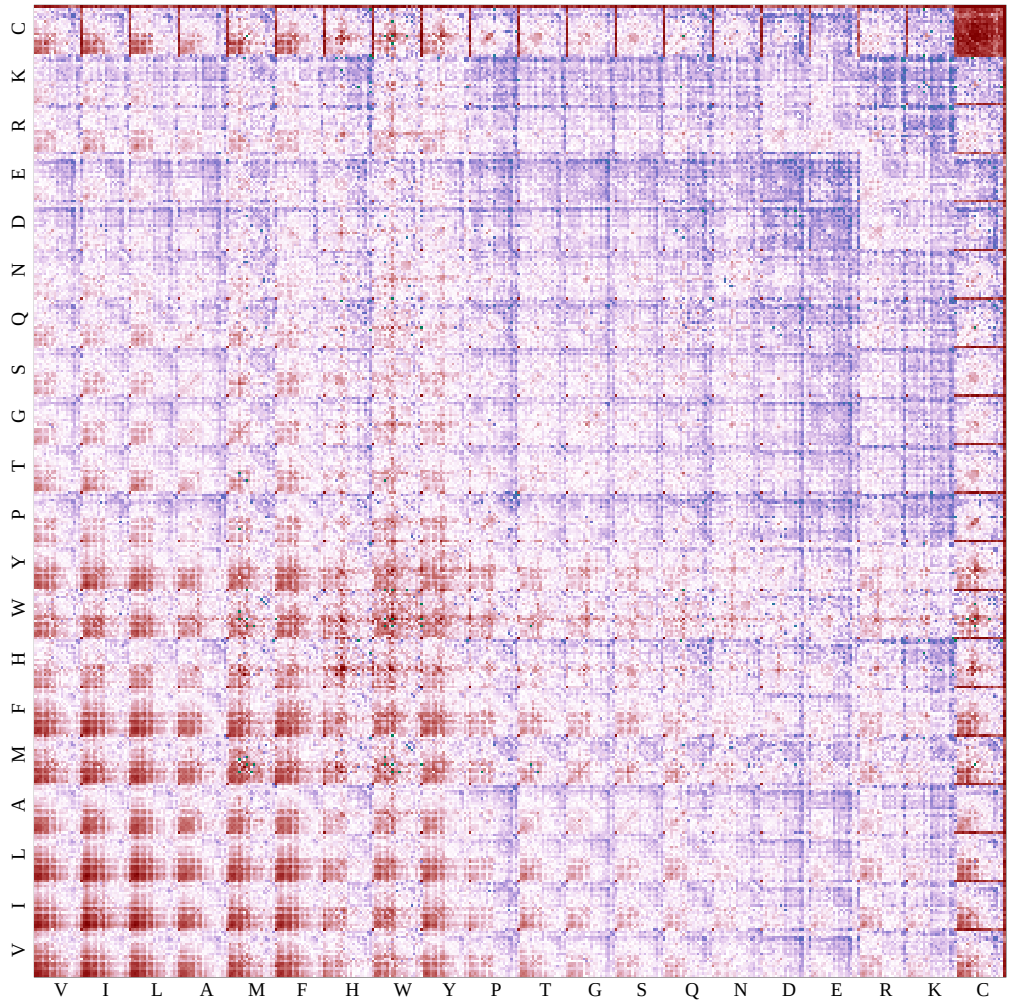


Figure 16: The 20 letter Tessellation potential, calculated for C^{α} atoms, using the filter procedure.

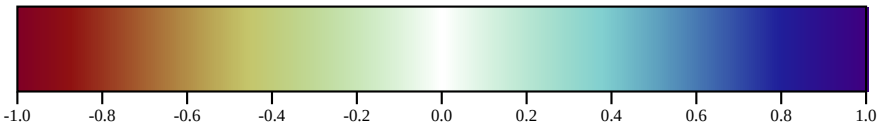


Figure 17: Color coding as used in the plot of the potential

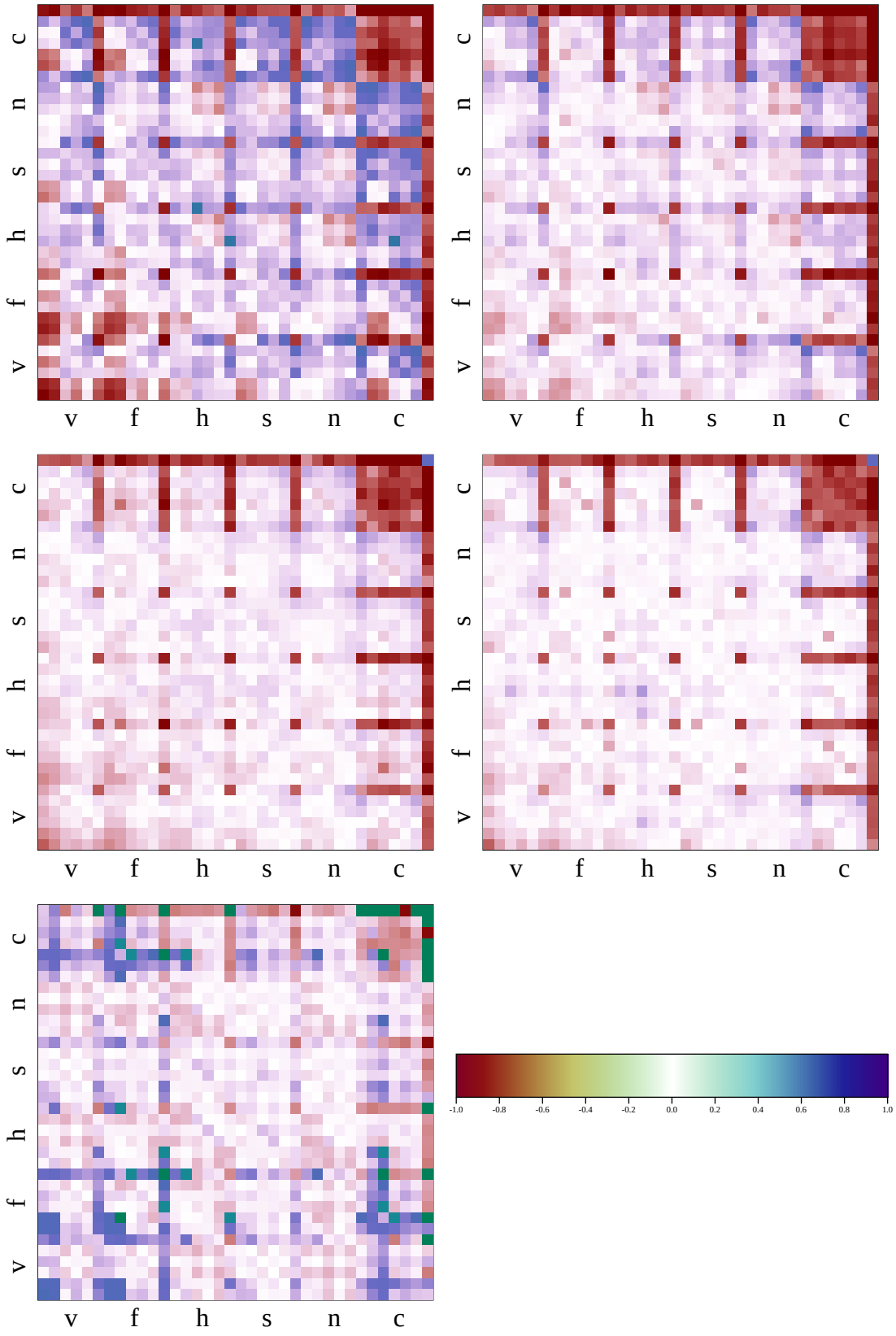


Figure 18: Graphical representation of the 6l potential as provided by A. Tropsha [68]

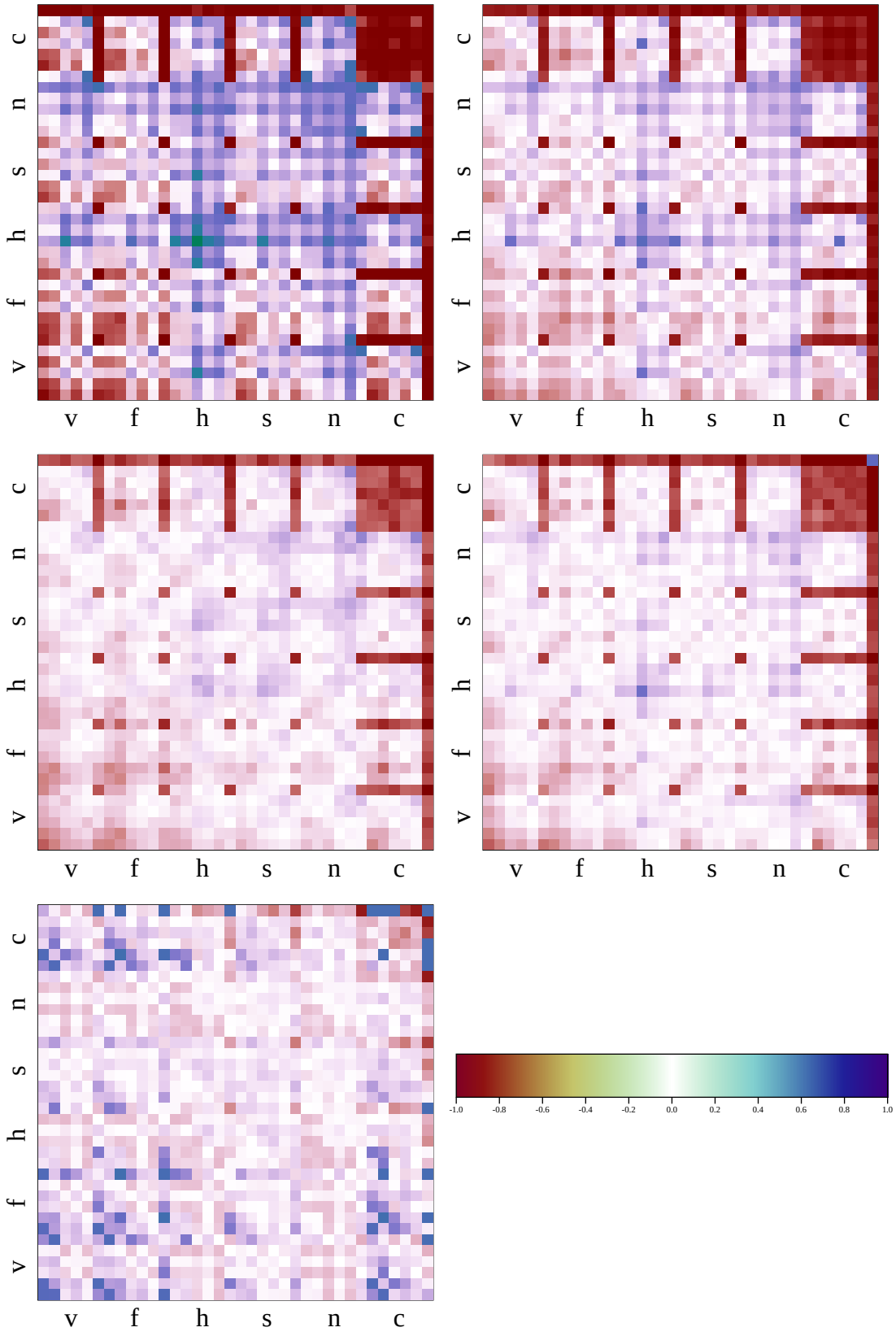


Figure 19: Graphical representation of the 6l encoded potential derived from the *pdb-select* dataset for C^{α} atoms with filter applied. Each Delauney class is shown in a seperate plot, starting with class 0 in the top left corner. All scores are normalized to fit the interval $-1 \leq z_{qilkj} \leq 1$.

4.3 Inverse-folding

4.3.1 Example: Thioredoxin

The example used in the inverse fold calculation was the `2trx` structure of Thioredoxin. The structure is represented with the superb resolution of 1.68Å from crystallographic data, the source was *Esterichia coli*. The asymmetric unit contains two molecules, named "`A`" and "`B`". It has been chosen due to the fact that it is a well-known globular structure. The secondary structure as calculated by `stride` [22] can be seen in figure 20.

Its biological function is electron carrier, it acts as electron donor in the reduction of ribonucleotides and plays an important role in the dark reaction of photosynthesis. It has regulatory capacity on other enzymes by reducing their disulfide bridges. The active form of Thioredoxin contains two cysteins, which are oxidized to form a disulfide bridge, when reducing other S-S bonds. It is reactivated by ferredoxin.

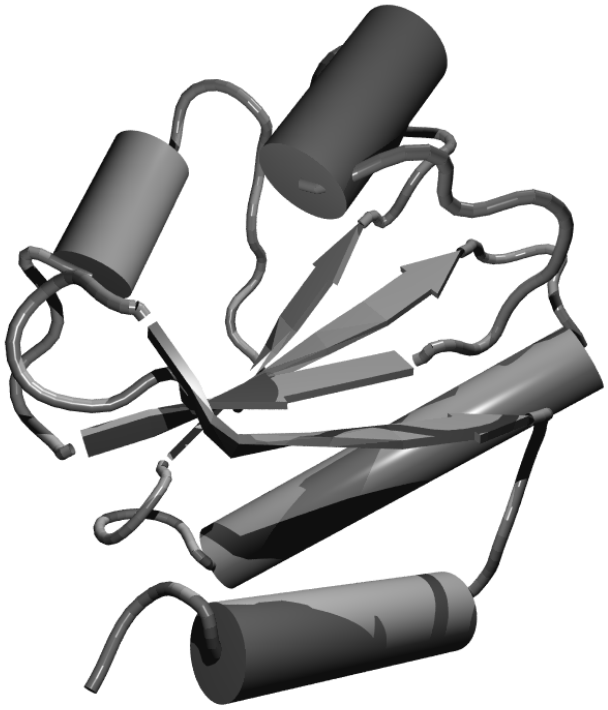


FIGURE 20: Secondary structure for the `A` molecule of Thioredoxin, as determined by `stride`. Sequence (108 amino acids): `SDKIIHLDDSFDTDVLKADGALVFWAEWCGPCKMIAPILDEIADE-YQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLA`

4.3.2 Previous results

Checking sequences that were optimized to fold into a particular structure by an adaptive walk using e.g. PROSA with the Tropsha potential showed a consistent progression. But the main back-draw in the statistical geometry approach was that if one took sequences that were optimized by the Tropsha potential and calculated scores for these sequences with other kinds of potentials (e.g. PROSA) they were considered as increasingly bad.

Viewing adaptive walks using the 6l potentials showed at least that PROSA accepts the sequences as getting better (figure: 21), but the score reached is far from native. For the 20l alphabet the situation is even worse, it seems that the scores go worse while optimizing (figure 22). Especially the PROSA surface scores were completely unacceptable bad. This was the observation, that gave rise to the use of a particular surface term. Much better is the acceptance of PROSA optimized sequences within the Tropsha potential (figure 23)

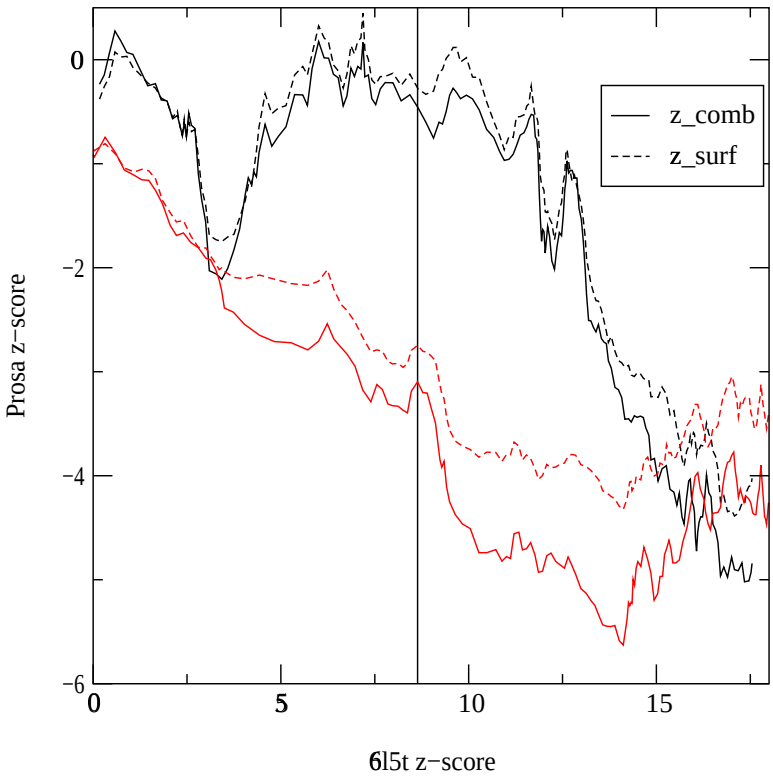


FIGURE 21: Data derived from the original potential as provided by A. Tropsha. The sequences were optimized, using the 6l5t Tropsha potential, cross check with PROSA. The filter was applied for tessellation, but not for calibration. The plot shows two runs, the surface score is shown as dashed line, the normal line represents the combined scores.

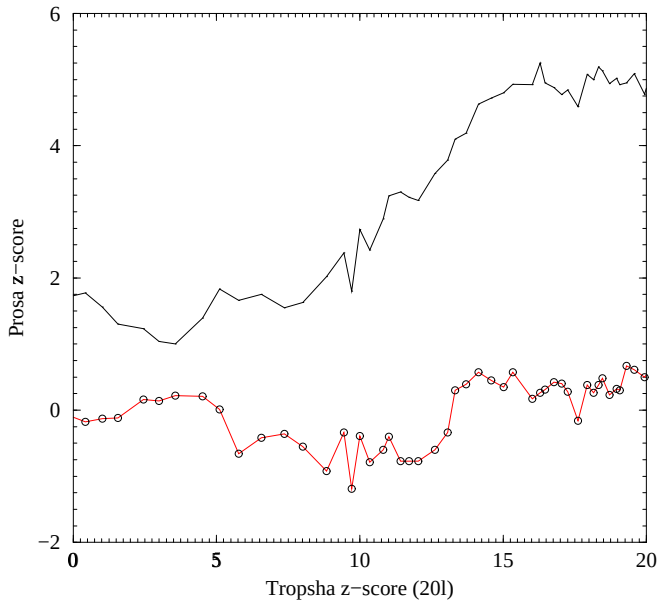


FIGURE 22: Adaptive walk using the original Tropsha potential. The lower curve shows the PROSA surface scores, the upper curve shows the combined scores. The sequences were optimized, using the 20l Tropsha potential afterwards cross check with PROSA. Filtering was applied for tessellation, but not for calibration. Note: positive PROSA scores indicate very bad z-score

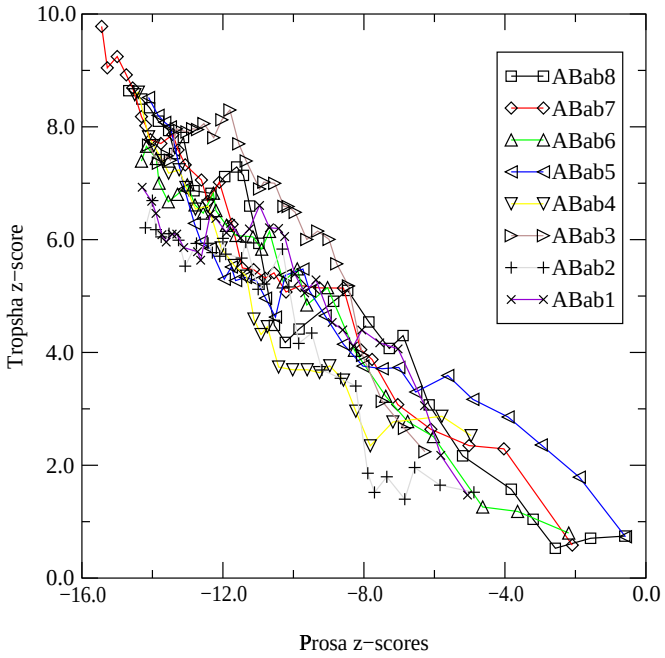


FIGURE 23: Tropsha scores for PROSA optimized sequences: The potential data used was the original Tropsha derived values. In both potentials C^{α} was used for calibration, PolyProtein: `poly10k.pdb`

4.3.3 Gaining Significance

Probing the potential in a "real life" situation means to perform adaptive walks, and check the resulting sequences in other force fields. An important criterion to test is whether a native like score can be reached, and figure 24 shows that the enhanced tessellation potential does this in a Hamming distance of about 50, what is less than halve the sequence length for `2trx`.

Figure 25 shows the same cross check of `PROSA` and tessellation sequences as described before. It can be seen, that adaptive walks using the enhanced tessellation potential now show increasing `PROSA` scores as well, though the `PROSA`-wild-type score of the structure is still not reached.

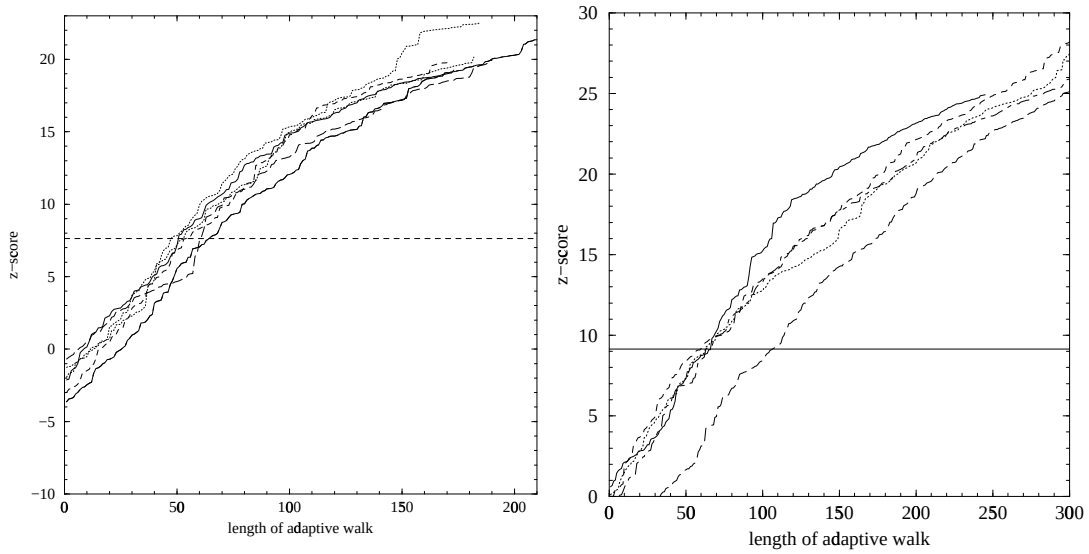


FIGURE 24: Adaptive walks performed for `2trxA` using the tessellation potential:
The left plot shows adaptive walks using the 6l potential, the horizontal dashed line at 7.8 marks the z-score of the native sequence. The adaptive walk was performed using C^{β} atoms, filtering, PolyProtein: `poly10k.pdb`
The right figure shows adaptiv walks using the 20l alphabet, the horizontal line at 9.1 marks the z-score of the native sequence. The walk was performed using the 20l potential, C^{β} atoms, filtering, PolyProtein: `poly10k.pdb`

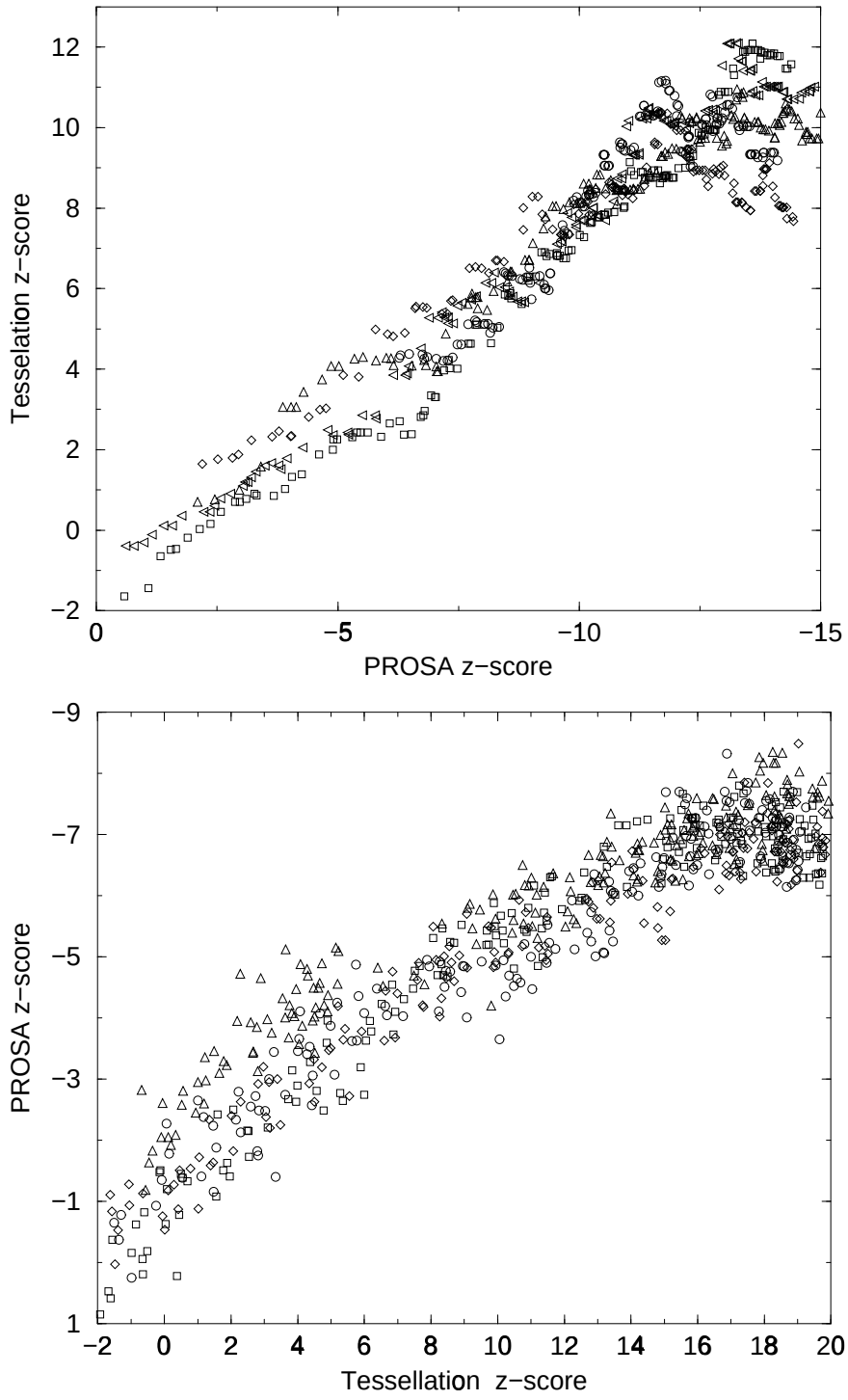


FIGURE 25: The upper figure shows adaptive walks performed using `PROSA` to inverse fold `2trxA`. C^{β} atoms and combined scores were used in both cases of z-score calculation. Using the `poly10k.pdb`-PolyProtein.
In the lower plot adaptive walks for `2trxA` using the tessellation potential are shown. The cross check was performed with `PROSA`. Conditions: C^{β} atoms, filter on, 20l alphabet,

5 Conclusion and Outlook

5.1 Summary

This work shows that extending the Tropsha four point potential provides significant improvement in consistency and accuracy. The major goals were to superpose a specific surface term and to implement C^{β} atoms for backbone representation. Checking the discriminative power of the potential as well as cross validation experiments showed an increase in reliability. A solid basis for future developments was laid by the implementation of the calibration tool. It is now easy to adapt the parameters to different coding schemes and to recalibrate the tessellation potential using revised versions of the *pdb-select* database. Applications that can use the parameter sets like `tropinverse` or `tropscore` are efficient enough to make large scale simulations feasible.

Critics on knowledge-based potentials as in [63] have to be judged carefully: The intention to gain access to "protein-like-energies" by lattice simulation is not very useful, since a 2-letter (**HP**) protein would not fold at all, since it's energy landscape is much too simple [66]. It has to be emphasized, that inverse folding does not rely on a "real" energy, but introduces the z-score as a relative scale. For the purpose of exploring a sequence-structure mapping this is satisfactory.

Nevertheless it is clear that modeling a protein without considering the amino acids as a chain is far from being accurate. The use of Delauney classes as shown for the 6 letter alphabet could help to consider chains. Also the effect of volume exclusion has to be added to improve the potential's quality. This work shows, that extending the Tropsha potential by biophysical necessary terms leads to improvements. The number of extensions can be increased easily if each contribution is kept as an additive term, just as the surface. Otherwise the number of parameters would explode and the statistical significance of the data would be lost. These additional "potential parameters" must be calibrated for the set as well, keeping the same conditions for all superposed terms. This is true due to the fact that depending probabilities may be factorialized, under respect of their dependency [4].

5.2 Directions for Future Improvements

The package could serve the scientific community beside `PROSA II` as mean to check experimental derived protein structures. Also if sequences of high homology to proteins with known structures are found by experiments, the inverse folding approach can give first hints for structural relationships.

To improve the discriminative power and comfort of the knowledge based potential it is planed to introduce some further terms:

- The only parameter in finding the neighboring set is the cut-off from applying the filter. This could be avoided by using a water shell, as described by Zimmer *et al.* [70]. The water molecules would be placed on a virtual grid around each residue, thereby fulfilling the constrain of a minimum distance to any neighboring residue. The water molecules will be treated as a further class of residues, having contact to exposed amino acids, and hence contribute to the overall statistics. Contacts with a certain "water content" could be excluded this way.
- Though the *pdb-select* database provided by Hobohm *et al.* [31] is well suited for statistics of sequences there are chains within the set that show abnormally low z-scores when the native sequence is threaded to its structure. In most cases this could be attributed to obvious reasons (e.g. membrane proteins, chains wrenched from the core, etc.). It would be preferrable however to exclude these abnormal chains already from the calibrating dataset. The calibration could be fully automated, as an iterative scheme provided a reasonable way of pre-processing the *pdb-select* can be implemented.
- Reduced alphabets showed an improvement in the statistics of the database. A careful selection of different coding schemes is expected to emphasize this effect.
- Once a good representation of a volume term is derived, it will be streighforward to use it as an extention of the potential.
- For the user's comfort a GUI will to be developed. The collection of all tools will be made available as package via the internet.

Appendix

A Programs

The efficient implementation of the algorithms described so far was the goal of this work. The availability of a decent tool to use *any* dataset for calibrating a potential force field opens a lot of possibilities for evolutionary and biophysical studies using energetic parameters. All programs were written in ANSI C for attaining maximum portability and speed. At the moment only the Linux, SGI and DEC-Alpha versions exist. An MS-DOS version is not planned, since this platform is not suited for a serious computing ☺.

A.1 Calibration of the Potentials

NAME

`calibrate` — calibrate the tessellation potential

SYNOPSIS

```
calibrate [-f {0,1}][-A {20,6}][-T {CA,CB}][-F list]
[-P path][-S]
```

DESCRIPTION

`calibrate` is a program that reads pdb-files specified in a list or from a path, performs a Delauney tessellation for each of the proteins and counts the occurrence of 4-tuples of residues. The log-likelihood of the quadruples printed to stdout.

OPTIONS

`-f [0,1]`	switch to turn on(1)/off(0) the filter criterion for irregular tetrahedra, default is on
`-S`	generate surface potential
`-T [CA,CB]`	use C^α or C^β atoms for calibration, default is C^α
`-r`	reject unsuitable chains and record these files in the output
`-A 20`	standard amino acid alphabet is used: default
`-A 6`	generate parameters for 6 letter alphabet
`-?`	display short usage message for the program

LIMITATIONS

If improper protein chains are input (e.g. chains with sequence gaps, rare amino acids etc.) the program stops unless the `-r` option was given in which case the files are rejected. Each rejected chain is noted in the output. Generally notes marked "`@`" are intended to be read by other applications (e.g. `tropinverse`).

Furthermore, the pdb files must fulfill some basic criteria: At least 6 atoms (either C^α or C^β) must be present, otherwise no useful tessellation is possible. The column order as outlined in the pdb-contents-guide has to be fulfilled, otherwise parsing is impossible. The only exception is the atom numbering of PolyProteins, since these numbers are greater than 10 000 (5 columns are read for this variable). For the construction of the virtual C^β of the bf G residues C^α, N and C coordinates are necessary.

For contact types with zero observations the calculated score will be $-\inf$. These values have to be replaced in a post-processing step, before using the potential with `tropscore` or `tropinverse`. Useful values for the replacements are 0 or the minimum score for other contact types of the class.

Ideally only globular soluble proteins without large ligands should be used for extracting the potential. Since `calibrate` cannot automatically recognize unsuitable protein chains, an appropriate selection should be prepared beforehand. A good strategy is to exclude all chains that have poor scores using some other existing potential.

PERFORMANCE

The hardware requirements are low: a maximum usage of 10 MB RAM was observed, processing of a database of 700 proteins takes

about 10 minutes on an i386 based machine (PentiumPro™ 200MHz), most time is spent in IO procedures. The program takes input from command the line, which makes it very easy to script the process. The detailed algorithm is shown in form of a flow chart in figure 26. The `qhull` package used for the tessellation can be found at: `www.geom.umn.edu/software/download/qhull.html`.

SAMPLE SESSION

Example: Generation of a 20 letter contact parameter set for the *pdb-select* release June 98, using a filter and C^{β} atoms for calibration:

```
~> calibrate -r -T CB -A 20 -f 1 -F list_pdb_J98
############################################################
#                    C A L I B R A T E                     #
#$Id: calibrate.c,v 1.16 1998/10/02 08:01:48 gw Exp gw $#
############################################################
# command given:calibrate -r -T CB -A 20 -f 1 -F list_pdb_J98
# pdb-file list from: ../data/dir_all_Jun_98
# STANDART DELAUNEY POTENTIAL MODE
@ potential for CB atom of the chains
@ filter for triangulation is ON
@ using 20 letter alphabet
############################################################
/scr/pdb/Jun98/119l.pdb_
/scr/pdb/Jun98/153l.pdb_
/scr/pdb/Jun98/1a0a.pdbA
...
```

Output :
The parameters as well as comment lines indicated by "#" or "@" are sent to stdout and should be redirected to a file. The list of processed files as well as warnings and error messages are written to stderr. Parameters for the 20 letter alphabet output as follows:

```
#cont    obs.freq    exp.freq  obs/ex           q  Obs.
AAAA   0.00013638  0.00004207    3.24  0.51081024    76
```

If the 6letter variant was chosen (option `-A 6`) the output of the parameters would be:

```
#cont  class 0  class 1  class 2  class 3  class 4          Obs.
cccc    4.3341   3.1592   0.0000  -0.5000   -0.800  102:19:1:0:0
```

The calculation of a surface potential file is as simple, the option `-S` is required additionally. After calibration a post-processing has to ensure that all terms are proper, i.e. the $-\inf$ for contacts which were not observed has to replaced by some estimate before feeding the data into `tropscore` or `tropinverse`.

A.2 Additional Tools

Before the *pdb-select* is parsed for the actual calibration process, a pre-formating of the heterogenous files is performed. A PERL script has been created for this task:

NAME

`backbonextract.pl` — a tool to process pdb files for later use in calibrate

SYNOPSIS

`backbonextract.pl` -in path

DESCRIPTION

`backbonextract.pl` is a PERL script to pre-process pdb-files, for later use in `calibrate`. It searches for files with extension `.pdb` in the directory `path` and extracts the backbone of protein chains to the current directory. Each chain is written into a separate file, by appending the chain identifier to the original pdb file name. A log containing a time stamp and information on each protein chain read is written to the file `logfile` (overwriting existing files of the same name).

OPTIONS

`-in`	directory to be searched for pdb files
`-? or --help`	display short usage for the script

LIMITATIONS

- PERL-Version 5 is required.
- nucleotides are omitted.
- theoretical models and files with nucleic acid content are skipped. Only the first model of several (NMR) models is used.
- Chains with less than 30 ATOMS are skipped.
- In case of alternate locations the 'A' or '1' version is taken
- Chains with gaps are discarded, identical chains are not detected
- The files **must** have the extension `.pdb`

The parameters obtained from `calibrate` can be conveniently visualized as shown in figures 16. Such PostScript™ plots are produced by `plot_trop`:

NAME

`plot_trop` — generate PostScript™ plots from a tessellation potential parameter file.

SYNOPSIS

`plot_trop` parameter_file

DESCRIPTION

`plot_trop` encodes the log-likelihoods from tesslellation potential parameter files in either grey-scale or color and writes PostScript™ output to stdout.

OPTIONS

`-? or --help`	display short usage for the script
`-6`	the specified parameter file is 6 letter encoded and uses five Delauney classes
`-c`	color mode (default is grey-scale).

LIMITATIONS

All log-likelihoods are scaled to fit the interval of $-1 \leq q_{ijkl} \leq +1$. Surface parameters can not (yet) be plotted in the current version.

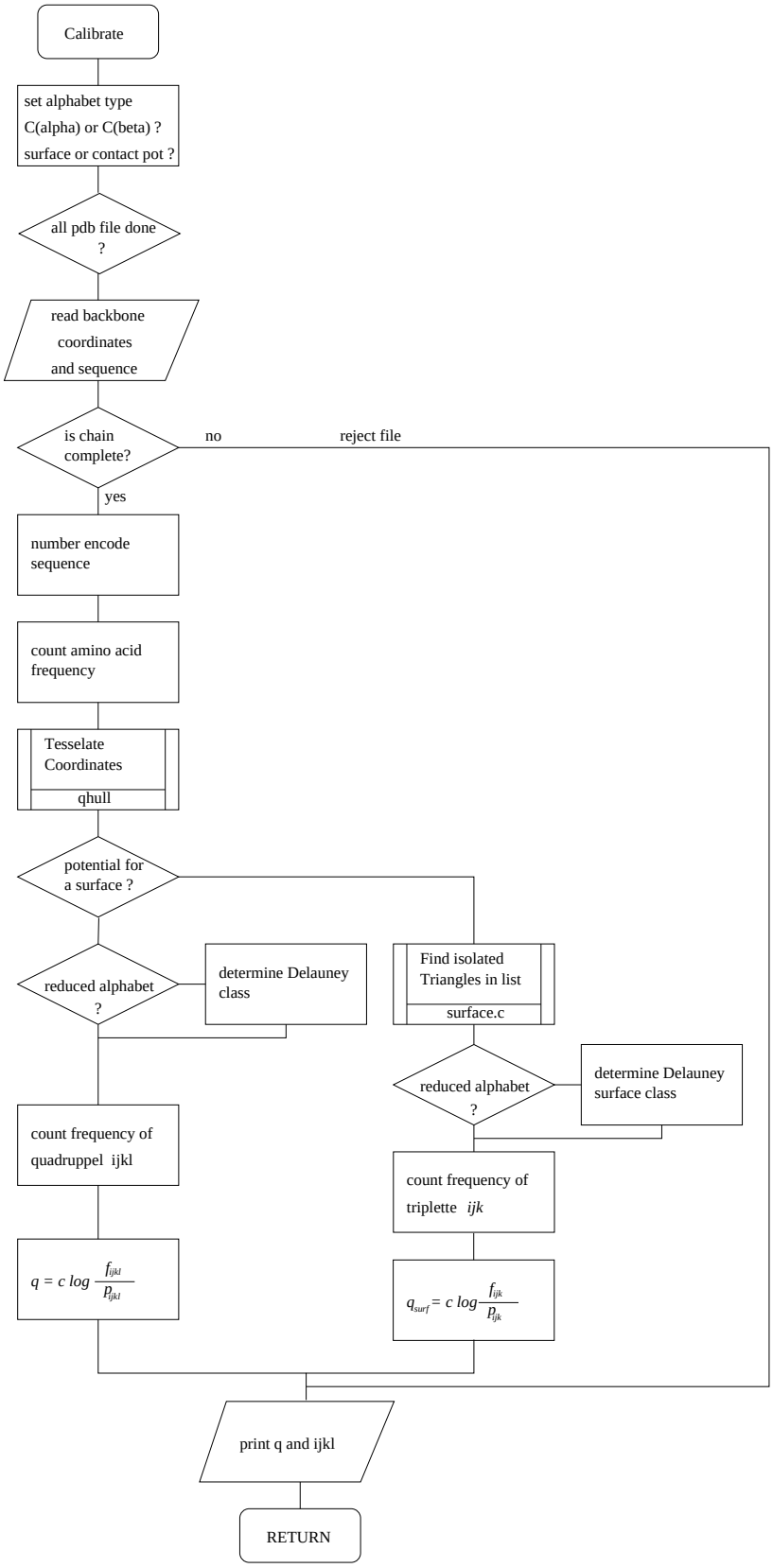


FIGURE 26: Flow chart of the algorithm implemented for the calibration tool for four point potentials. For tessellating the `qhull` algorithm is used. See page 55 for command line parameters

A.3 Tessellation z-score Calculation and Inverse Folding

Routines for calculating z-scores using the potentials created by `calibrate` have been implemented in C. Currently, there are two basic applications that make use of these routines and potentials. Of course the same restrictions for pdb files as in in `calibrate` apply to these applications.

The `tropscore` program calculates z-scores for one or more sequences on a given structure, using contact potentials, surface potentials or both. Since the tessellation has to be performed only once scores for many sequences can be calculated in relatively short time.

NAME

`tropscore` — calculate z-scores using the tessellation potential

SYNOPSIS

```
tropscore [ -P pot_file and/or -S surf_pot ]
target.pdb polyprotein[.pdb or .tpp]
```

DESCRIPTION

The `tropscore` tool uses basically the same routines as tropinverse, and is thought as tool to calculate z-scores for a list of sequences (e.g. for cross checking experiments, using sequences optimized with other knowledge based potentials).

The Program calculates the z-score for a sequence if threaded to the structure of target.pdb. A PolyProtein is used as library for structures for the calculation. If it has the extension "`.tpp`", it is considered to be a saved tessellation from a previous run. After the tessellation (and surface generation if chosen) is proceeded, the program pauses for input of sequences from stdin. The z-score, energy and sequence used for calculation are written to stdout. In combined mode all three scores (contact, surface and combined) are calculated and printed.

OPTIONS

`-? or --help`	display short usage for the script
`-P`	tessellation potential parameter file, if no surface parameter file is given, the contact mode is chosen, otherwise combined potential mode
`-S`	surface parameter file, if no contact parameter file is used, it the surface mode is used
`-s` name.tpp	save tessellated PolyProtein to name.tpp
`-A 6 or 20`	alphabet to use, default: 20 letter
`-f 0 or 1`	switch filter on (1) or off (0), default: on
`-T CA or CB`	atom type to be used for calculation
`-F` factor	real number used as factor to combine surface and contact potential terms in combined mode

LIMITATIONS

- target and the sequence specified have to have the same length, gaps are not allowed.
- Lines in the parameter files starting with "#" or "@" are ignored
- pdb-files have to fullfil the same conditions as for `calibrate`

The `tropinverse` program designs sequences with good z-scores on a predefined structure, thereby solving the inverse folding problem. Sequences are optimized using adaptive walks with the z-score as fitness function. The evaluation of scores under the chosen conditions employs the same routines as in `tropscore`. Again the tessellation has to be performed only once. It is recommendable to save the result of the PolyProtein tessellation to disk if more than one run is intended. Start sequences for the walk can be either provided from stdin or are generated randomly by `tropinverse`. An iterative process of mutation and score evaluation follows, stopping only if the best sequence is found or a certain number of trials is exceeded. For details of the procedure see the flowchart in figure 27.

NAME

`tropinverse` — inverse folding using the tessellation potential

SYNOPSIS

```
tropinverse [options] [-P cont.pot or/and -S surf.pot]
target.pdb polyprotein.pdb
```

DESCRIPTION

`tropinverse` is an implementation of an adaptive walk algorithm for inverse folding of protein sequences, using the tessellation four point potential. It uses z-scores as fitness criterion to find sequences that are likely to fold into a given structure. The start sequence can either be given from stdin, or is generated by the program. Sequence, scores and Energies for the individual steps of the walk are printed to stdout.

OPTIONS

`-d` dump the program to disk

`-e` start sequence from random amino acids, given the "environment" class as defined by Eisenberg [9]

`-n` start sequence from random amino acids, using the mean frequencies as in the Swiss Prot database:

A	C	D	E	F
.0760	.0176	.0529	.0628	.0401
G	H	I	K	L
.0695	.0224	.0561	.0584	.0922
M	N	P	Q	R
.0236	.0448	.0500	.0403	.0523
S	T	V	W	Y
.0715	.0581	.0652	.0128	.0321

`-r` start sequence is purely random

`-s` start sequence using the mean amino acid frequency as in the Swiss Prot, propensities as in Chou-Fassmann [16].

`-? or --help` display short usage for the script

`-P` tressellation potential parameter file, if no surface parameter file is given, the contact mode is chosen, otherwise combined potential mode

`-S` surface parameter file, if no contact parameter file is used, it the surface mode is used

`-s` name.tpp save tessellated PolyProtein to name.tpp

`-A 6 or 20` alphabet to use, default: 20 letter

`-f 0 or 1` switch filter on (1) or off (0), default: on

`-T CA or CB` atom type to be used for calculation

`-F` factor real number used as factor to combine surface and contact potential terms in combined mode

`-?` display short usage

LIMITATIONS

- It is of course impossible to translate a reduced alphabet back to the 20 letter. If the given sequence is 6l encoded, a random amino acids is related to the class.
- The parameter files (contact or surface) must have the format of the `calibrate` output, again lines starting with "#" or "@" are ignored

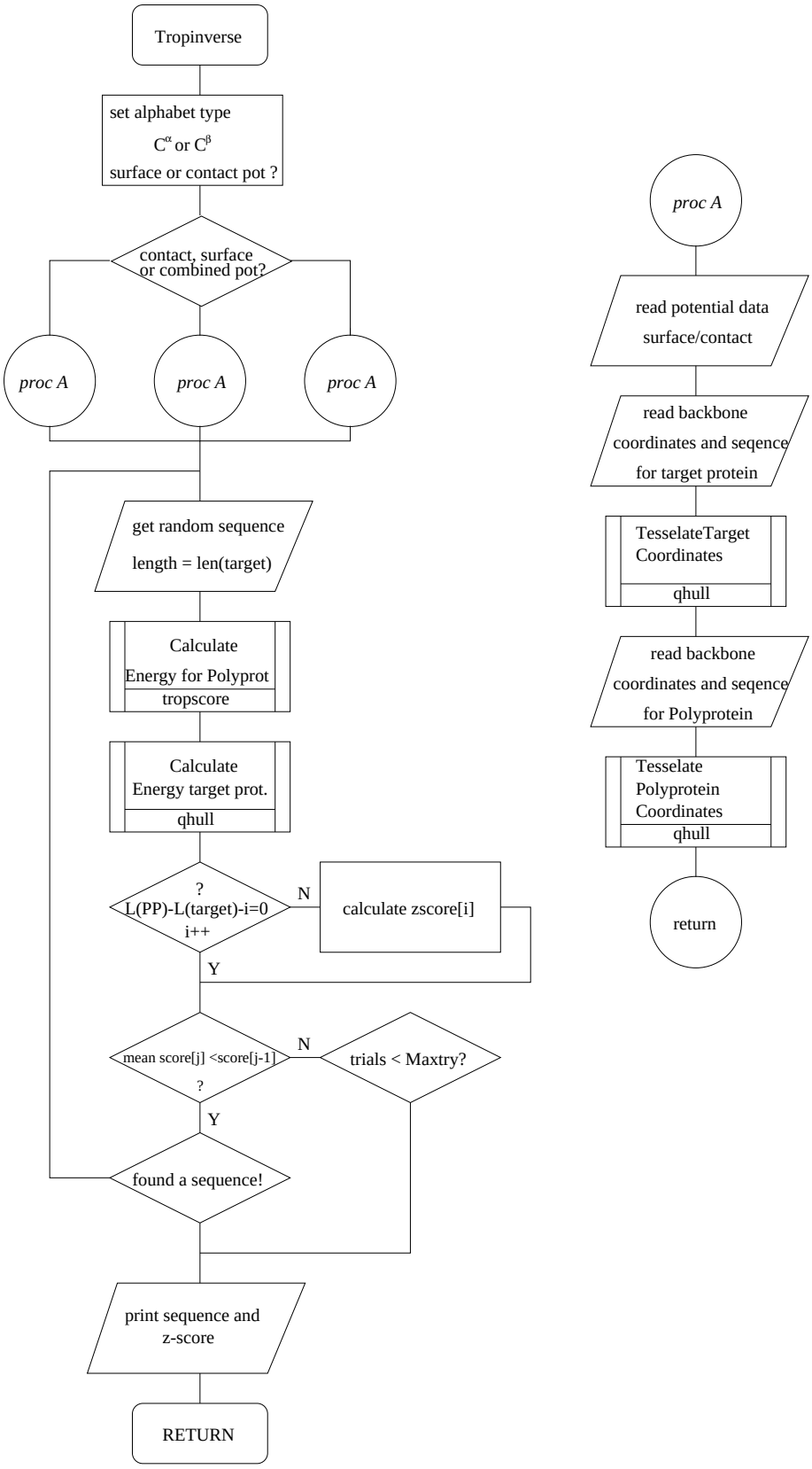


FIGURE 27: Algorithm for inverse folding of proteins using the tessellation potential as used in `tropinverse`

B Abbrevations

Å	Ångstrøm (1 Å= 10^{-10}m)
DNA:	Desoxyribonucleic Acid
EMBL:	European Molecular Biology Laboratory
ftp:	file transfere protocoll
HP:	Hydrophobic - Polar
MD:	Molecular Dynamics
PDB:	Protein Data Bank
PROSA:	Protein Structure Analysis
RNA:	Ribonucleic Acid
SAW:	Self-Avoiding Walk
URL:	Universal Resource Locator

C List of Tables

D List of Figures

E PDB Select

These are the IDs of the 900 files provided as *pdb-select* release in June 1998. The last letter represents the polypeptide chain.

1UXD_	1CFE_	1CFH_	1CFO_	1ULP_	1CDS_	1BVH_	1BW3_	1VIG_	1C5A_
1VHP_	1CDB_	1TUM_	1CTL_	1CTO_	1CUR_	1TFB_	1DDF_	1DEC_	1TIH_
1TSG_	1CMR_	1TPT_	1TPM_	1TLE_	1TIT_	1VTX_	1ZTO_	1AZ6_	1BAK_
1BBO_	1ZDD_	1BCN_	1ZWD_	2BDS_	1AWJ_	2AT2A	2ADX_	2ABD_	1AYJ_
1BCT_	1BNB_	1BOR_	1WKT_	1WIU_	1BTB_	1VVC_	1WTUA	1ZAQ_	1YUB_
1BFMA	1BGK_	1BHB_	1BIP_	1DEF_	1NOE_	1HSN_	1NKL_	1NGR_	1NFA_
1IFE_	1HRYA	1OCP_	1HCD_	1HEV_	1NRE_	1HMCB	1HQI_	1MSEC	1JVR_
1LEFA	1LEB_	1KUL_	1KSR_	1KRT_	1JLI_	1MAK_	1IRL_	1IRSA	1ITF_
1IVA_	1IYV_	1GRX_	1RTNA	1EHS_	1ERD_	1ROO_	1ROF_	1EXH_	1SHCA
1DEG_	1SVQ_	1SRO_	1DPI_	1EAL_	1SKYE	1FBR_	1PIH_	1PFT_	1PFSA
1PDC_	1PCE_	1GPT_	1FWP_	1FDM_	1RES_	1QYP_	1PUT_	1PRR_	1POU_
1TIV_	1AQS_	2FOW_	2FSP_	2HP8_	2IL6_	1APJ_	1APF_	2VGH_	1ARK_
3DPA_	1AG2_	2EZH_	1AG4_	2VIK_	2EZK_	1AP8_	2RGF_	1AJYA	2PTL_
2NCM_	2NEF_	2PLDA	2PAC_	1AH9_	2KTX_	1AOY_	1AHK_	2TBD_	2STWA
1AJ3_	1AAF_	1AT7_	7GATA	1ACP_	1AA3_	2CDX_	1ATY_	4RNPA	5ZNF_
1AB3_	2BI6H	2ECH_	1AFP_	1CMYB	3R1RA	2LDB_	1MYPA	1GYLA	1AONO
1IF1B	1GNHA	1YSTH	4HMGA	1RRF_	1SMVC	1FPKA	1ASYA	1AGNA	2MEV4
2MEV1	1KCW_	1PLR_	1CRKA	1AIPE	1RPT_	2BPA2	1PDGC	2BPA1	1CNE_
1NOM_	4RHV1	1TAHA	1PYP_	1GLEF	2UCZ_	1VBA4	1FZAB	1FZAA	1GUKA
1CMKE	1DHX_	1FC1A	1HWHA	2BCT_	1BCFA	1RUSA	1TFPA	1DKTA	1SERA
1PKN_	1BLE_	1CDI_	1STD_	1PYSA	4DPVZ	1FRVA	1FRVB	1BMFG	1SQC_
1TNRA	1AVOB	1JRHI	1NFDA	2LGSA	1THJA	1CID_	1FGJA	1AOAA	1DKGB
1ASX_	2BBVA	1PKP_	1RGS_	1AGRE	2VAOA	1BCMB	1FOKA	1HLOA	1LFB_
1ATNA	1EBPA	1XXCA	1PREA	1DLHA	1OCCE	1OCCK	1OCCH	1OCCC	1OCCG
1OCCF	1OCCD	2DRPA	1MXA_	1PIOA	1GIN_	1QAPA	1SLY_	1TLK_	1NPOC
2DMR_	1LIAA	1BTMA	1BIB_	1YSC_	1ATIB	1IBCB	1IBCA	1CSGA	1BCPF
1BCPB	1D66A	1AR1A	1ANV_	1OFGA	1ZID_	1LXTA	1ECRA	1PEX_	1SMEA
1UDII	3PBGA	1CD1A	1AHJA	1AHJB	1GGTA	1AKJD	1XDTR	1HSTA	1AO7B
1AOI_	1JKW_	1KMMB	1AB8A	1AQIA	1SIG_	1DIV_	1LRV_	2EMO_	1FDZA
1HCNB	1HCNA	1LKTA	1PS1A	1CBY_	1LXA_	1BVP1	1LNH_	1LL1_	1AI6A
1P04A	1FCDA	1WSYB	1YTFC	1PDNC	1ITBB	1HJP_	1KZUB	1EXNB	1RHOC
1CKNA	1HTMB	1IPSA	1KB5B	1XBRB	1AUA_	1GTRA	1DUBB	1AIHA	1UMUB
1VPFA	1AN9A	1HMY_	1MSPB	1IHFA	1POIB	1POIA	1VDC_	2POLA	1AURA
1KNYA	8ATCB	1FUIA	1HJRA	1HLB_	1PYAB	2STV_	1LPBA	1TC3C	1JACA
1TDX_	1RLW_	1BP1_	1AFRA	1DAR_	1HULA	1JMCA	1LGHA	1RNL_	1UBY_
1CNT2	1PYDA	3ULLA	1VDEB	2MPRA	2TRCP	1BNCB	1CHKA	1YTW_	1ECEA
2LIV_	2MTAC	1NBBA	3MDDA	1PAX_	2OMF_	2PFKD	1FKX_	4AAHA	1STFI
1AK4C	1PYTA	1DJXA	1AK5_	1AQT_	2DYNA	2MASA	1GTQA	1GGGA	1ITG_
2RSLC	1AROB	1TABI	1INP_	1AERB	1CYX_	1SMPI	1PRCC	1JSUC	4PGMB
1YCQA	1OPR_	1DHY_	1BNDA	1CFYA	4HTCI	1IPWB	1KIT_	1AM7A	1AORA
1DHR_	1MNMC	1IGNA	1FT1A	1FT1B	1TIID	1QUF_	1MHLC	1JXPA	1GPC_
1ZXQ_	1NOYA	1BEO_	1SMTB	1AIJS	1AAO_	1GRJ_	1GCB_	1YCSB	1TUPC
1ETPA	2CAE_	1NSGB	1XVAA	1ECMB	1DRU_	1MAZ_	2BGU_	1LBA_	1BHMB
1KTE_	1VMOA	1DELB	1FTPA	1HTP_	1TUL_	2PHLA	1BOVA	1GNWA	1GTMA
1GPMB	1JLYA	1TDTC	1HCGB	1GOH_	1DHS_	1LPN_	5EAU_	1AYM3	1AYM2
1AWCB	1AUVB	1ANU_	1CFR_	1ASH_	1ABRB	1AUK_	1THTB	1A4SA	1ACC_
1SFE_	1P38_	1RMD_	1PEA_	1VNC_	1AS4B	1IRK_	1CSBB	1AXIB	1GPL_
1AX4A	1RYT_	1PTA_	2TCT_	1OTGA	2DKB_	1RLAA	1FJMA	1VHRA	1DEAA
1EFVA	1EFVB	1SMNA	2HHMA	1ESC_	1TFR_	1AOB_	1BROA	1VDFA	3PCHM
3MINB	3MINA	1AIKC	1AOL_	1ALY_	1WPOB	1VIN_	1KINB	1FMTB	1HAVA
2LBD_	1GOTB	1GOTG	1CSN_	1DPGA	1SPUA	1BTN_	1BFTA	1CFB_	1PBN_
1BBPA	1AF7_	1MSC_	1ECPA	1HOE_	1BV1_	1TAFA	1TAFB	1CEWI	2PGD_
2PTD_	1FJLB	4MT2_	1SRA_	1VID_	1WDCA	1DKZA	1R69_	1POC_	1AOCA
1GSA_	1JPC_	1NBAB	1LKI_	5CSMA	1LCT_	1SVPA	1TSP_	2SCPA	1ALO_
1WHTB	1NSYA	1RMG_	1GKY_	1PII_	2I1B_	1OYC_	1AQOA	1APYB	1DORA
1FWCA	2HPDA	1PNE_	1OBPA	1MKAA	1RVAA	2FIVA	1AA2_	1TRKA	2RSPB
1GARB	1OSPO	2PSPA	1PTQ_	1AQ6A	1LTSA	1FURB	1OVAB	1DAAA	1URNA

1FUA_	1RSS_	1OIS_	1HSBA	1ECL_	1ATO_	1SKZ_	1NEU_	1MLDA	1STMA
1MAI_	1CLC_	1ESFA	1AD2_	1AGQD	1AOZA	1OTFA	1YASA	1EDT_	1OPY_
7AHLA	1FLEI	1DXY_	1JDW_	1LATB	1NCIB	1AK1_	1CEO_	2CHSA	2KINA
2KINB	1SFTA	1BGP_	1REC_	1PNKB	1GPB_	3BCL_	1IDK_	1LIS_	1RSY_
1ISO_	2FHA_	1SVB_	1WHO_	1CPO_	1AIL_	1REGY	1KVU_	1CHMA	1VPSA
1HGXB	3TSS_	2TYSA	1GIFA	1SLTA	1EDE_	1ADOA	1DUPA	1YAIA	1BYB_
2PII_	1ONRA	1LML_	1FIT_	1VLS_	1DOKA	1PUD_	2ABK_	1PTY_	1MUCA
1ZNBA	1QBA_	1TIB_	2HTS_	1EUR_	1GND_	1UCH_	1HCRA	1V39_	1NFN_
1AKO_	1AH6_	2SPCA	1UXY_	1VCAA	1YTBA	1LCL_	1POT_	1MSK_	1MML_
1QNF_	1NPK_	1AYL_	1MPGA	1NBCA	1SLUA	1NULA	1TIF_	1WBA_	1BDO_
1LMB4	1IIBA	1AFWA	2POR_	1UAE_	1TML_	1TYS_	1BEBA	2SAK_	1FNA_
1UNKA	1GDOB	1AL3_	1GD1O	2SICI	1AOQA	1HXN_	1VIF_	1GUQB	1ATZB
2TGI_	1KVEA	1KVEB	2ACY_	1BDMA	1PGS_	1UBI_	1LBU_	1AMP_	1ATLA
1NAR_	3COX_	1CYDA	2NACA	2BAA_	1AXN_	1MZM_	2VHBA	1BRNL	1PDA_
1HA1_	1TVXA	1XGSA	1CHD_	6GSVA	1KPTA	2BBKL	2BBKH	1THV_	3CLA_
1XJO_	1PCFA	1AQZB	1AM3_	1BGC_	1AJJ_	1TADA	1TFE_	1XIKB	1KAZ_
2CYP_	1IDO_	2BOPA	1VJS_	4PGAA	1FVKA	1IDAA	1WAB_	1VHH_	6CEL_
1PDO_	1AGJA	1PMI_	1FDR_	1MTYD	1MTYB	1MTYG	1KID_	1ONC_	1SBP_
1FDS_	1THX_	1MOLA	2MSBB	1LT5D	1AKO_	1KNB_	2GDM_	1VSD_	1GAI_
2HFT_	1MWE_	2CCYA	1ANF_	1DOSA	2HMZA	3CHY_	5HPGA	1ERV_	1YVEI
1AQB_	1PHNB	1CNV_	119L_	1CEM_	1PHP_	1HTRP	1BTKB	1WER_	1VCC_
8RUCI	1GVP_	1ADS_	1AAYA	1JER_	2DRI_	1EDG_	1PHC_	1BKF_	1SMD_
1DAD_	5NUL_	1AOP_	3NUL_	1NWPA	1MRP_	1ARV_	3CYR_	1NIF_	1MRJ_
1ZIN_	1AJSA	1LAM_	2SIL_	1KUH_	1CSH_	2ILK_	1PPN_	1BFG_	153L_
3PTE_	2AYH_	4XIS_	1NOX_	1AWSA	1AKZ_	1HFC_	1LIT_	1AVMA	1RA9_
1TCA_	3GRS_	1ORC_	2CBA_	1KPF_	1AH7_	5ICB_	1AIE_	1WHI_	1RIE_
1OPD_	1LUCB	2ARCB	1EZM_	1CKAA	1ISUA	2MCM_	3B5C_	1EDMB	1POA_
2HBG_	2ENG_	2SNS_	8ABP_	1XNB_	1XSOA	2RN2_	1G3P_	1VWLD	1ABA_
2END_	1RPO_	1MBD_	2PHY_	1ECA_	1XYZA	1SGPI	256BA	2CTC_	3SDHA
1BENB	1AWD_	1RCF_	1NXB_	1PPT_	1RHS_	5P21_	1UTG_	5PTP_	1PLC_
1AAC_	1FUS_	1RRO_	1JHGA	7RSA_	1MSI_	1WEA_	1YCC_	1JETA	2PTH_
2SN3_	1AMM_	1CSEI	1CSEE	1ARB_	1IFC_	1RGEA	1IGD_	1CTJ_	5PTI_
8RXNA	1LKKA	2ERL_	1CEX_	1IXH_	1AHO_	1NLS_	2FDN_	3LZT_	1CBN_

F References

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[2] A. Babajide, R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes, and P. F. Stadler. Exploring protein sequence space using knowledge based potentials. *Protein Science*, 1998. submitted, Santa Fe Institute preprint 98-11-103.

[3] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 2:261–269, 1997. Santa Fe Institute Preprint 96-12-085.

[4] R. T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lodon*, 53:370–418, 1763.

[5] J. D. Bernal. *Nature*, 183:141, 1959.

[6] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[7] T. J. Blundell, T. L. Sibanda, J. E. Sternberg, M, and J. M. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326:347–352, 1987.

[8] J. U. Bowie, N. D. Clarke, and C. O. Pabo. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, 7:257, 1990.

[9] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–169, 1991.

[10] J. U. Bowie, R. Lüthy, and D. Eisenberg. Assesment of protein models with three-dimensional profiles. *Nature*, 356:83–85, 1992.

[11] C. Bradford, Barber, D. P. Dobkin, and H. T. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22:469–421, 1996.
`URL: http://www.acm.org`.

[12] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways and the energy landscape of protein folding: A synthesis. *Proteins*, 21:187–195, 1995.

[13] G. Casari and M. J. Sippl. Structure-derived hydrophobic potential: Hydrophobic potentials derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, 224:725–732, 1992.

[14] H. S. Chan and K. A. Dill. The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.*, 92,5:3118–3135, Mar. 1990.

[15] H. S. Chan and K. A. Dill. Comparing folding codes for proteins and polymers. *Proteins*, 24:335–344, 1996.

[16] P. Y. Chou and G. D. Fasman. Prediction of protein conformation. *Biochemistry*, 13:222–245, 1974.

[17] G. M. Crippen. Prediction of protein folding from amino acid sequences of discrete conformation spaces. *Biochemistry*, 30:4232–4237, 1991.

[18] G. M. Crippen and Y. Z. Ohkubo. Statistical mechanics of protein folding by exhaustive enumeration. *Proteins*, 32:425–437, 1998.

[19] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yeo, P. D. Thomas, and H. S. Chan. Principles of protein folding: a perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[20] A. V. Finkelstein, A. Y. Badretdinov, and A. M. Gutin. Why do protein architectures have Boltzmann-like statistics? *Proteins*, 23:142–150, 1995.

[21] M. S. Friedrichs and P. G. Wolynes. Toward protein tertiary structure recognition by means of associative memory Hamiltonians. *Science*, 246:371–373, 1989.

[22] D. Frishman and P. Argos. Knowledge-based secondary structure assignment. *Proteins*, 23:566–579, 1995.

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

[24] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci.*, 89:9029–9033, 1992.

[25] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. Optimized energy functions for tertiary structure prediction and recognition. In H.Bohr and S.Brunak, editors, *Protein Structure by Distance Analysis*. IOS Press, 1994.

[26] T. Grossman, R. Farber, and A. Lapedes. Neural net representations of empirical protein potentials. *Ismb*, 3:154–61, 1995.

[27] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monatsh. Chem.*, 127:355–374, 1996.

[28] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monatsh. Chem.*, 127:375–389, 1996.

[29] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. Identification of native protein folds amongst a large number of incorrect models — the calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.*, 216:167–180, 1990.

[30] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3:52–524, 1994.

[31] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein data sets. *Protein Science*, 1:409–417, 1992.

[32] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[33] L. Holm and C. Sander. Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.*, 26:316–319, 1998.

[34] J. J. Hopfield. Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proc. Natl. Acad. Sci.*, 84:8429–8433, 1987.

[35] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *J.Mol.Graph.*, 14:33–8, 1996.

[36] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci.*, 93:397–401, 1996.

[37] K. K. Koretke, Z. A. Luthey-Schulten, and P. G. Wolynes. Self-consistantly optimized statistical mechanical energy functions for sequence structure alignment. *Protein Science*, 5:1043–1059, 1996.

[38] J.-B. d. M. Lamarck. *Philosophie Zoologique*. A. Kröner, 1909. Deutsch von H. Schmidt Leibzig.

[39] C. Levinthal. Are there pathways for protein folding? *J. Chem. Phys.*, 65:44–45, 1968.

[40] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669, 1996.

[41] P. J. Munson and R. K. Singh. Statistical significance of hierarchical multibody potentials based on delauney tessellation and their application in sequence-structure alignment. *Protein Science*, 6:1467–1481, 1997.

[42] A. G. Murzin. New protein folds. *Curr. Opin. Struct. Biol.*, 4:441–449, 1994.

[43] A. G. Murzin. Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.*, 6:386–394, 1996.

[44] A. Neumaier, S. Dallwig, W. Huyer, and H. Schichl. New techniques for the construction of residue potentials for protein folding. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and R. D. Skeel, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes Comput. Sci. Eng.* Springer, Berlin, 1999.

[45] K. Nishikawa and T. Noguchi. Predicting protein secondary structure based on amino acid sequence. *Methods in Enzymology*, 202:31–44, 1991.

[46] C. A. Orengo, D. T. Jones, and J. M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372:631–634, 1994.

[47] B. Rost and C. Sander. Jury returns on structure prediction. *Nature*, 360:540, 1992.

[48] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70 % accuracy. *J. Mol. Biol.*, 232:584–599, 1993.

[49] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[50] E. I. Shakhnovich, V. Abkevich, and O. Ptitsyn. Conserved residues and the mechanism of protein folding. *Nature*, 379:96–98, 1996.

[51] E. I. Shakhnovich and A. M. Gutin. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci.*, 90:7195–7199, 1993.

[52] E. I. Shaknovich. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett*, 72:3907–3910, 1994.

[53] E. I. Shaknovich and A. M. Gutin. A new approach to the design of stable proteins. *Protein Engineering*, 6:793–800, 1993.

[54] R. K. Singh, A. Tropsha, and I. I. Vaisman. Delauney tessellation of proteins: Four body nearest neighbor propensity of amino acid residues. *J. Comp. Biol.*, 3:213–221, 1996.

[55] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force — An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[56] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *Journal of Computer-Aided Molecular Design*, 7:473–501, 1993.

[57] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[58] M. J. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5:229–235, 1995.

[59] M. J. Sippl, M. Ortner, M. Jaritz, P. Lackner, and F. Hannes. Helmholz free energies of atom pair interactions in proteins. *Folding & Design*, 1:289–298, 1996.

[60] J. M. Smith. Natural selection and the concept of protein space. *Nature*, 225:563–564, 1970.

[61] M. E. Sternberg. Secondary structure prediction. *Curr. Opin. Struct. Biol.*, 2:237–241, 1992.

[62] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.*, 25:115–130, 1996.

[63] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.*, 257:457–469, 1996.

[64] A. Šali and T. L. Blundell. Comparative protein modeling by satisfaction of spatial restrains. *J. Mol. Biol.*, 234:779–815, 1993.

[65] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chemistry*, 7:230, 1986.

[66] P. G. Wolynes. As simple as can be? *Nature Structural Biology*, 11:871–874, 1997.

[67] L. Zhang and J. Skolnick. What should the Z-score of native protein structures be ? *Protein Science*, 7:1201–1207, 1998.

[68] W. Zheng, S. J. Cho, I. I. Vaisman, and A. Tropsha. Statistical geometry analysis of proteins: implications for inverted structure prediction. In L. Hunter and T. Klein, editors, *Biocomputing: Proceedings of the 1996 Pacific Symposium*, pages 614–23. World Scientific Publishing Co, 1996.

[69] W. Zheng, S. J. Cho, I. I. Vaisman, and A. Tropsha. A new approach to protein fold recognition based on delaunay tessellation of protein structure. In L. Hunter and T. Klein, editors, *Biocomputing: Proceedings of the 1997 Pacific Symposium*, pages 486–97. World Scientific Publishing Co, 1997.

[70] R. Zimmer, M. Wöhler, and R. Thiele. New scoring schemes for protein fold recognition based on voronoi contacts. *Bioinformatics*, 14:295–308, 1998.

[71] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[72] M. Zuker. The use of dynamic programming algorithms in RNA secondary structure prediction. In M. S. Waterman, editor, *Mathematical Methods for DNA Sequences*, pages 159–184. CRC Press, 1989.

G Curriculum vitae

Full Name:	Günther Karl Weberndorfer
Place and date of birth:	31.05.1972 in Steyr,Upper Austria
Education:	Volksschule Reichraming Bundesgymnasium Steyr HTBLA Wels Chemische Betriebstechnik
Matura:	June 1991, "*Mit Auszeichnung*"
Military service	Oct.1991 - May 1992 NBC-defence Austrian army
Studies:	1992 - 1999 Biochemistry, University of Vienna
Address:	Institut für Theoretische Chemie Währingerstr 18 A - 1090 Wien gw@tbi.univie.ac.at