Writing with amino acids: designing the folding and binding of model proteins
Coluzza, I.

Citation for published version (APA):
Coluzza, I. (2005). Writing with amino acids: designing the folding and binding of model proteins
2 Artificial Evolution of a Lattice Heteropolymer

The typical size of single-domain protein is of the order of $10^3$ atoms (see Fig. 2.1). This number of atoms does not include the surrounding solvent that is, however, an essential element of the protein activity. Upon inclusion of the first shell of water the number of atoms increases by a factor ten.

Using the fastest computers available, it is now becoming possible to perform fully atomistic simulations of the folding of small proteins. However, as explained in the introduction, our aim is different: we wish to explore the constraints placed upon protein design by the requirements of foldability and function. For such studies it is advisable to use a highly simplified protein model. In what follows, we describe proteins as heteropolymers “living” on a 3D cubic lattice.

2.1 General properties of heteropolymers on a lattice

This lattice-polymer model for proteins is highly simplified. First of all, the side chain of the single amino acids are not taken into account, leaving the protein as necklace of beads with isotropic interactions. The major effect of this approximation is to ignore the entropic contribution of the side chain and to reduce the effect of steric hindrance. The second approximation consists in constraining the residues of the chain on a cubic lattice of unit side length. Constraining a polymer to a lattice does not change the topology of the chain, but all internal degrees of the monomeric units, as well as the vibrations of the bonds between them, are ignored. The model assumes that there are only nearest-neighbor interactions between the amino acids. The total configurational energy of a particular sequence in a given structure is given by

$$E = \sum C_{ij} S_{ij},$$

(2.1)

where $i$ and $j$ are particle indices, $C$ is the contact map defined as

$$C = \begin{cases} 1 & \text{if } i \text{ neighbor of } j \\ 0 & \text{otherwise} \end{cases},$$

(2.2)

and $S$ is the interaction matrix. For $S$ we use the 20 by 20 matrix fitted by Miyazawa and Jernigan [3]. From the PDB database these authors extracted the frequency of contacts for each pair of amino acids in a wide range of different proteins. The formation of a bond in
2 Artificial Evolution of a Lattice Heteropolymer

Figure 2.1: Hen egg white lysozyme is a single chain of 129 residues (~1000 atoms). It has an alpha+beta fold, consisting of five to seven alpha helices and a three-stranded antiparallel beta sheet. The enzyme is approximately ellipsoidal in shape, with a large cleft in one side forming the active site. Lysozyme on of the better known single domain proteins.
the collapsed polymer can be schematically represented by the following chemical reaction,

\((\sigma - \ast) + (\pi - \ast) \Rightarrow (\sigma - \pi) + (\ast - \ast)\) where \((\sigma - \ast)\) is the average binding energy between an amino acids of type \(\sigma\) and all the other types of amino acids. Using the quasi-chemical approximation Miyazawa and Jernigan estimated the free energy associated with the formation of a bond between all possible pairs of amino acids. It should be stressed that it is not at all obvious that the frequency of contacts between amino-acids is a direct reflection of their binding strength. For one thing, it is obvious that the "training" set from which the interaction energies were deduced is a subset (and probably not a random subset) of all existing protein structures. Note that here and in the next chapters the energy is expressed in units of \(kT\) relative to the energies in the interaction matrix.

A given lattice polymer can form a large number of compact conformations. Obviously, every conformation is characterized by a different contact map. Hence, the energy of the polymer depends on its conformation. In a mean field approximation the energy spectrum of the compact structures of a random chain on a lattice has the shape shown in Fig. 2.2.

The mean field expression for the entropy is [4, 5, 6]

\[ S(E) = \begin{cases} 
N \ln \gamma - \frac{E^2}{2N\sigma_B^2} & \text{if } E > E_c \\
0 & \text{if } E \leq E_c
\end{cases} \quad (2.3) \]

\(^1\) where \(N\) is the number of elements in the chain, \(\sigma_B\) is the standard deviation of the interaction matrix, and \(\gamma\) is the coordination number for fully compact structures on the lattice. \(E_c\) is the

\(^1\)In the definition of the entropy the contribution of the quantity \(\sqrt{\pi \sigma_B^2}\) is ignored, as explained by Derrida [6]
Artificial Evolution of a Lattice Heteropolymer

(lower) crossing point of the parabola with the abscissa, \( E_c = -N \sigma B (2 \ln \gamma)^{1/2} \). The finite width of this energy spectrum reflects the fact that the system is frustrated. The "native state" corresponds to the least frustrated structure. If the native state is non-degenerate, this lowest-energy conformation has zero entropy. The degree of frustration of a heteropolymer is linked to the number of different monomers that it contains. This is particularly obvious in the case of a homopolymer. For such molecules, all compact states are unfrustrated and have the same energy. This picture is confirmed in our simulation, where indeed we observe a non-degenerate native state for a well designed sequence. In the following we will refer to the lowest energy state as the native state of the heteropolymer.

In 1993 Shaknovich and Gutiin [7, 8, 9] showed that it is possible to "design" a lattice protein in such a way that it will fold into a specific conformation. They achieved this by optimizing the sequence of amino acids, using a Monte Carlo algorithm that randomly exchanges amino acids within the chain molecule. The acceptance of such trial swaps depends on the energy change associated with the move

\[
\Delta E = \sum (S'_ij - S_{ij}) C_{ij}
\]

(2.4)

where \( S' \) (\( S \)) denotes the interaction matrix of the new (old) sequence of amino acids. During a Monte Carlo run of several million cycles, a large number of distinct sequences are generated. The sequence \( S^* \) with the lowest energy is assumed to be the best candidate to fold into the native state.

\[
E_{\text{Native}} = \sum C_{ij} S'_{ij}.
\]

(2.5)

A closely related, but different method, for the design of heteropolymers that fold into a specific structure, is the so called "Painted globule" model (see, e.g. [13, 14]). The central idea behind this approach is to look at the target structure, and then distinguish between surface and core residues (hydrophilic and hydrophobic). The design consists of a sequence of folding and re-painting steps.

In our work we have chosen to not use this last method, because it does not allow for a complete control on the target configuration. The annealing process needs equilibration steps of the configuration of the chain. In practice this means that the minimization procedure is done on both the \( C \) and the \( S \) terms of the energy. To study designed configurational changes in protein we found it crucial to have the target structure equal to the final native state of the protein.

Pande et al. [10, 11, 12] have provided a theoretical analysis of the general design of a foldable protein sequence. In particular, these authors showed that, in the context of the Random Energy Model, the phase behavior of designed protein sequences can be predicted analytically. One of the main findings of Pande et al. was that the energy gap separating the target "native" state from the set of non-native compact states, is inversely proportional to the design temperature (the fictitious temperature at which we perform Boltzmann sampling of different sequences for a given target conformation). This a crucial result because it gives a theoretical basis to the feasibility of heteropolymer design, and also because it fixes a thermo-dynamic relation between the process of folding of a chain and the the design of its sequence of monomers. In the next section we will give the derivation of such a fundamental result.
2.2 Folding and Freezing Transition

In this section we derive the relation between the freezing transition for heteropolymers with random sequences and the folding transition of designed proteins. Although, the derivation it is valid only in a mean-field approximation, the final result will give a clear and simple physical explanation of what it means to design a protein.

In their 1997 review Pande et al. [11], presented a new approach to describe the statistical mechanics of protein folding. The approach that we follow in subsequent chapters has been inspired by the methods described in ref. [11].

Let us start by considering the total free energy of a random heteropolymer

\[ \mathcal{F}(T) = \langle \mathcal{F}_{\text{seq}}(T) \rangle = -T \langle \ln Z_{\text{seq}}(T) \rangle \]  

(2.6)

where \( \mathcal{F}_{\text{seq}}(Z_{\text{seq}}) \) is the free energy (partition function) for a possible random sequence and the average is done on all possible realizations. As we discussed above, the entropy in REM (Eq. 2.3) vanishes for states with energy below \( E_c \), while above it the average density of states \( \langle n(E) \rangle \) is equal to \( \mathcal{M} P(E) dE \) where \( \mathcal{M} = \gamma^N \) is the number of states, \( dE \) in the thermodynamic limit can be approximated by \( dE \approx \sqrt{N} \gamma^2 \), and \( P \) is the Gaussian density probability

\[ P(E) = \frac{1}{(2\pi\sigma_B^2)^{1/2}} \exp \left[ -\frac{E^2}{2\sigma_B^2} \right] \]  

(2.7)

where \( \sigma_B^2 \) is the variance of the interaction matrix. We can then rewrite the partition function in a simpler form by taking advantage of the exponential weight of states in the saddle point \( E^* \) of the density of state

\[ Z_{\text{seq}}(T) = \int n(E) e^{-E/T} dE = \int \mathcal{M} P(E) e^{-E/T} dE \approx \mathcal{M} P(E^*) e^{-E^*/T} \]  

where \( E^* = -N\sigma_B^2/T \) as the value of the energy at which the argument inside the integral is maximum for a Gaussian distribution \(^3\). This representation is valid for temperatures for which \( E^* \) is located in the continuous part of the spectrum. In this regime, the partition function is independent of the particular sequence and in Eq. 2.6 we can swap the ordering of taking the average and the logarithm. The main assumption of the REM is that each interaction between the monomers is independent, and follows a Gaussian distribution. Using this approximation we can compute the average of \( Z \) over the all possible sequences, as the product of the contribution of each pair interaction:

\[ \langle Z \rangle = \mathcal{M} \left[ \int P(E_{\text{pair}}) e^{-E_{\text{pair}}/T} dE_{\text{pair}} \right]^L = \mathcal{M} e^L \left[ E - \frac{\sigma_B^2}{2T} \right]^L \]  

(2.8)

\[ \mathcal{F} = -T \ln \langle Z \rangle = L \left[ E - \frac{\sigma_B^2}{2T} \right] - T N \ln \gamma = L \left[ \frac{E}{T} - \frac{\sigma_B^2}{2T} \left( 1 + \frac{T^2}{T_c^2} \right) \right] \]  

(2.9)

The approximation for \( dE \) is an arbitrary power \( \alpha \) of \( N \), with \( \alpha < 1 \) Ref. [6]

\(^3\)Note that by taking the thermodynamic limit of the log of the density of states \( \ln \gamma P(E) dE \approx \gamma^2 P(E) \sqrt{N} \approx N \left[ \ln \gamma - \frac{E}{2\sigma_B^2} \right] \) we recover the expression of the entropy in Eq. 2.3
where $L$ is the (independent of the conformation) number of contacts between monomer in any particular compact conformation, and $\bar{E}$ is the average of the interaction matrix. We expressed the right-hand side as a function of the transition temperature $T_g$ at which the entropy $S(T) = -\frac{dF}{dT}$ vanishes

$$T_g = \sigma B \sqrt{\frac{L}{2N \ln \gamma}} \quad (2.10)$$

This temperature is called glass temperature because below it the system is trapped in one of the conformations that belong to the discrete region of the density of states. We now have the statistical tools to describe the phase behavior of a quenched random heteropolymer. The reason why we sketch the derivation of Eq. 2.9 explicitly is that it illustrates the peculiar temperature dependence of the system. Above the glass temperature $T_g$, the random-energy heteropolymer explores many states practically independent of the particular sequence of amino acids. However, as the temperature is lowered, the equilibrium is dominated by few discrete states of low energy, which are highly dependent on the particular sequence. The transition at $T = T_g$ is called the freezing transition [15, 16]. Initially it was suggested that the random-energy model might provide a useful model for protein folding, as it yields a unique ground state with a probability independent of the system size. However the energy differences between structurally distinct states in the discrete region of the energy spectrum are only of the order of $\sqrt{N}$, which does not allow for a robust equilibrium state. The question is then if it possible to design particular sequences that freeze into a robust ground state. In order for such an approach to work, the energy of the target state must be well separated from the boundaries of the continuous distribution of states, where the glassy states accumulate (at typical distances of order $\sqrt{N}$). Using mean-field arguments similar to the ones used above, we can derive an expression for the average energy of the designed state $E_d$ as function of the temperature of the canonical ensemble of sequences $T_d$. We start by choosing a target conformation $C_d$ as our tentative native state. This conformation is characterized by an energy $E_d = \mathcal{H}(S_d,C_d)$ that obviously depends on the sequence $S_d$. The partition function obtained by summing over all possible sequences is denoted by $W$ and it defines a free energy $F_W$ through

$$F_W \equiv -T_d \ln W(T_d) = -T_d \ln \left( \left< \exp\left[ -\mathcal{H}(S_d,C_d) / T_d \right] \right> \right)$$

$$\approx \left< \mathcal{H} \right> - 1/2T_d \left[ \left< \mathcal{H}\mathcal{H} \right> - \left( \left< \mathcal{H} \right> \right)^2 \right]$$

$$= L \left[ \bar{E} - \frac{\sigma B^2}{2T_d} \right].$$

where $T_d$ denotes the design temperature. In terms of $F_W$ we can write an approximate expression for the average energy of the designed sequence $\left< E_d \right> = -\frac{\partial \ln W}{\partial (1/T)} \bigg|_{T \rightarrow T_d}$, which does not depend on the target conformation, but instead show that the energy is inversely proportional to the design temperature

$$\left< E_d \right> = L \left[ \bar{E} - \frac{\sigma B^2}{T_d} \right]. \quad (2.11)$$

This result implies that if the design procedure is carried out at a temperature lower than $T_g$, then the average energy of the designed state will be below the boundaries of the continuous
part of the density of states. The lower the design temperature, the larger the gap will be and, as a consequence, the designed state will be increasingly robust. Furthermore the freezing transition of sequences designed at low \( T_d \) will occur at a temperature higher than \( T_g \) because the energy of a designed sequence is lower than the one of the glassy states of a random heteropolymer. The folding temperature \( T_f \) is defined as the temperature at which there is equilibrium between the native (target) state and the random globule. \( T_f \) can be computed by comparing Eq. 2.11 with Eq. 2.9

\[
\frac{1}{T_f^2} + \frac{1}{T_g^2} = \frac{2}{T_f T_d}
\]

(2.12)

which is independent of the mean value of the interaction matrix, but it does depend on the variance \( \sigma_B \) (see Eq. 2.10). We expect that the rate at which a designed protein folds into the native state is faster than the rate at which a completely random heteropolymer reaches its lowest-energy state. The reason is that folding takes places at a temperature \( T_f > T_g \) where the system has enough thermal energy to overcome local energy barriers. Using the relation in Eq. 2.12, we constructed a phase diagram that describes the general relation between design and folding in heteropolymers (see Fig. 2.3).

### 2.3 Design algorithm

In the previous section I have presented a mean field analysis of the design process of heteropolymers. The results proved the validity of a general design scheme, that would simply minimize the total energy of the polymer, with a fixed target compact configuration in a canonical ensemble of sequences. If this minimization process was carried out at a temperature \( T_d \) low enough, than the final heteropolymer would have shown a folding behavior similar to real proteins. I now will describe the technical details our implementation of the design algorithm. In order to design monomer sequences that yield a target conformation, we used a modified version of the Shakhnovich method. We perform point mutation of single amino acids and swap of identity between two randomly chosen residues. Unlike the latter method, our approach does not keep the amino acid composition of a chain fixed. Rather, we allow for random changes of amino acids. As a consequence we had to devise a criteria more sophisticated than the normal metropolis scheme to prevent that this compositional sampling results in the formation of homopolymers of the most attractive amino acid. We introduce a (purely fictitious) compositional “temperature”. Increasing the compositional temperature increases the compositional entropy. To perform the sampling, we combine the following acceptance criterion with the normal acceptance Metropolis rule

\[
P_{acc} = \min \left\{ 1, \left( \frac{N_{new}^P}{N_{old}^P} \right)^{T_p} \right\}.
\]

where \( T_p \) is the arbitrary parameter that plays the role of a temperature, and \( N_P \) is the number of permutations that are possible for a given set of amino acids. \( N_P \) is given by the multinomial
2 Artificial Evolution of a Lattice Heteropolymer

Figure 2.3: Phase diagram of the freezing transition in globular heteropolymers with a designed sequence. We have indicated with on the $y$ axis we plot the temperature $T$ of the system during the folding process, and on the $x$ axis there is the temperature $T_d$ of the design process. We can identify 3 phases: 1) a Glassy phase in the region $T_d/T_g > 1$ and $T/T_g < 1$, in which the protein is stuck in one of the low energy states in the discrete part of the energy spectrum of a random heteropolymer. 2) a Random phase for $T/T_g > 1$ if $T_d/T_g > 1$ and $T > T_f$, where $T_f$ is the folding temperature and is calculated through Eq. 2.12 with the conditions $T_d/T_g < 1$ and $T_f/T_g > 1$. This region is called Random because the system fluctuates between the different conformation of the continuous part of the density of states, and it correspond to the unfolded state of a design protein. 3) The last region is the Folded phase where the target conformation dominates the equilibrium. The thermodynamic stability of this phase is directly dependent on how much the design temperature $T_d$ is lower than the glass one $T_g$. 

![Phase diagram of the freezing transition in globular heteropolymers](image-url)
expression

\[ N_p = \frac{N!}{n_1!n_2!n_3!...} \]  

(2.13)

where \( N \) is the total number of monomers and \( n_1, n_2 \) etc are the number of amino acids of type 1,2,\ldots. With this condition we can explore a large set of sequences, yet avoid the formation of homopolymers. In the absence of any \textit{a priori} criterion to fix \( T_p \), we used trial and error. If \( T_p \) is too small, the chains will tend to become homopolymers (with a degenerate native state). In contrast, when \( T_p \) is too large, we noticed that the lowest energy sequences were no longer able to fold (the sequence becomes too random). We should therefore choose a value of \( T_p \) that yields a compromise between these two conflicting tendencies. To explore a range of values for \( T_p \) and at the same time limit the trapping in local minima of sequence space we introduced a parallel tempering algorithm for the sequence sampling at different pseudo-temperatures. We use the following criterion to decide whether the swapping of the sequences between two adjacent temperatures would be accepted:

\[ P_{\text{acc}} = \min \left\{ 1, \left( \frac{N_p^{\text{new}}}{N_p^{\text{old}}} \right)^{\Delta T_p} \right\} \]  

(2.14)

In our simulations we found that good sequences could be designed if parallel tempering was performed with the set of values \( T_p = \{1, 1/2, \ldots, 1/14\} \). With this set of compositional temperatures we obtained native states that were both stable and non-degenerate. We stress that the fictitious temperature parameter \( T_p \) only plays a role in the generation of suitable sequence. It plays no role in the subsequent simulations of chain (re)folding. We used variations of the above design algorithm for each individual problem that we studied. In every chapter we describe the modifications introduced and we explain the reasons for introducing the approach used in that chapter. Of course, once we have generated a particular sequence (or two, as the case may be), we still need to test whether these sequences do indeed fold into the desired structures.