ELSEVIER

Biophysical Chemistry 105 (2003) 105–118

Biophysical Chemistry

www.elsevier.com/locate/bpc

A simple procedure to weight empirical potentials in a fitness function so as to optimize its performance in ab initio protein-folding problem

Luigi Agostini[1], Stefano Morosetti*

Department of Chemistry, University of Rome 'La Sapienza', P.le A. Moro 5, Rome I-00185, Italy

Received 19 February 2003; received in revised form 14 April 2003; accepted 14 April 2003

Abstract

In an approach to the protein folding problem by a Genetic Algorithm, the fitness function plays a critical role. Empirical potentials are generally used to build the fitness function, and they must be weighted to obtain a valuable one. The weights are generally found by the comparison with a set of misfolded structures (decoys), but a dependence of the obtained fitness generally arises on the used decoys. Here we describe a general procedure to find out, from a given set of potentials, their better linear combination that could either identify the wild structure or prove their powerlessness. We use topological considerations over the hyperspace of the potentials, and a multiple linear inequalities solver. The iterated method flows through the following steps: it determines a direction in the hyperspace of the potentials, which identifies the native structure as a vertex among a set of misfolded decoys. A multiple linear inequalities solver obtains the direction. A Genetic Algorithm, tailored to the specific problem, uses the fitness function defined by this direction and generally reaches a new structure better than the experimental one, which is added to the ensemble. The decoys so generated are not dependent on a deterministic criterion. This iterative procedure can be stopped either by identifying an effective fitness function or by proving the impossibility of its achievement. In order to test the method under the hardest conditions, we choose numerous and heterogeneous quantities as components of the fitness function. This method could be a useful tool for the scientific community in order to test any fitness proposed and to recognize the most important components on which it is built.

Keywords: Genetic algorithm; Linear programming problem; Dynamic generation of the decoys; Principal component analysis (PCA); Haar transform; Principle of minimal frustration

*Corresponding author. Tel.: +39-06-49913730; fax: +39-06-4453827.
E-mail address: stefano.morosetti@uniroma1.it (S. Morosetti).
[1] Department of Inorganic Chemistry, University of Newcastle upon Tyne, NE1 7RU, Newcastle upon Tyne, UK.

0301-4622/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0301-4622(03)00130-3

1. Introduction

The determination of the folded protein conformation from amino acid sequence information by computational methodologies comes up against two main problems: to define a fitness function able to identify the native structure among alternative structures (decoys), and to explore a very wide conformational space, where the values of the fitness function form a landscape, which is generally rich in local minima.

As for this last problem, energy landscapes look like rugged surfaces with many deep valleys corresponding to local minima, and it has been suggested for a protein to be kinetically foldable, there must be sufficient overall slope so that the numerous valleys flow in a funnel [1,2] toward the native structure. Energy landscapes with competing interactions (which are responsible for the valleys) are termed 'frustrated', and a principle of minimal frustration [3] has been suggested for kinetically foldable proteins.

However, when the quantities forming a fitness function replace the real energies the computational energy landscape becomes in fact an 'artificial one', with a shape depending on the potentials used, besides the representation. Probably, it will change its appearance, eventually lacking the funnel form. In this case, the computational search will stay easily trapped in local minima.

Therefore, in order to reduce the trapping, one can conveniently act upon the several topics involved in the searching procedure, that is, the structure representation, the algorithm, the choice of the quantities used and their combination in a fitness function. We act on all these aspects except for the choice of the quantities, in order to develop an iterative procedure aimed to identify an effective fitness function (if it exists) from a set of given quantities.

1.1. The protein representation code

Modifying the representation has an important effect in changing the appearance of the energy landscape, as there is a topological alteration in the space in which the structures are embedded. Our choice is the discrete Haar transform of the backbone dihedral angles sequence [4], which is able to separate the local regular structure from the tertiary one and, therefore, seems to mimic to some extent the route the foldable proteins follow [5–7].

1.2. The search algorithm

We use a Genetic Algorithm with operators tailored to the peculiar problem, principally for its ability to handle high-dimensional objective functions with multiple local extrema.

Using a fitness function made up of empirical potentials and ad hoc criteria gives rise to the problem of combining several incommensurable quantities. There is no general theory known for the proper weighting of each fitness component. An infinite number of combinations of weights arises even for a small number of fitness components. Different authors [8–12] have found the weight values through many different simulation trials.

We limit ourselves to a linear combination of the quantities, so that the goal of the procedure will be to determine the values for the coefficients (weights) of the quantities setting up the fitness function. The method could be generalized to more complex functional forms, at the expense of a major number of parameters are to be determined, using an increasing number of terms of an orthogonal expansion (i.e. using the Legendre polynomials) for each quantity.

1.3. A topological description of the search strategy

An effective fitness function has a maximum (or minimum) value for the native structure, compared to any other structure. If we report all the structures in a hyperspace, where the axes are the quantities setting up the fitness, this property means that a hyperplane including the native structure leaves all the other points representing the non-native structures on the same side. These points can then be considered as belonging to a convex set, the native structure being an extreme point of the set. A straight line perpendicular to the hyperplane is an effective search direction for

the native structure, as the projection of the native representative point on it is an extreme in respect to the projections of all the other points. Therefore, the components of the line in the hyperspace represent the weights of the quantities that build the fitness function. Generally, there will be a polyhedral cone of solutions [13]. Scaling the axes does not affect these considerations because of the mathematical property, which states that a reversible application on a convex set leaves the extreme points unchanged [14].

Among the polyhedral cone of solutions, it is convenient to choose the one which maximizes the gap between the native structure and the misfolded ones. This will have the effect of choosing a landscape more similar to a funnel, and, therefore, reducing the trapping.

1.4. The choice of the misfolded structures

The choice of the misfolded structures (decoys) is a critical one. Generating decoys by a mechanistic method (i.e. the gapless threading) will not be in general representative of the whole conformational space [15]. Therefore, the fitness function calculated on the base of the described topological considerations will be dependent on the way in which decoys are obtained [16]. This problem becomes more serious when the fitness is utilized in a search algorithm, as virtually the complete space can be explored, and the algorithm must discriminate native folds not only from other fully collapsed structures, but also from expanded ones with correct secondary structures, from collapsed structures with good phase separation between hydrophobic and hydrophilic residues, etc. To avoid such problems, we use a 'dynamic' generation of the decoys, that is, they are generated during the searching steps of the procedure to identify an effective fitness function. In fact, the components of the fitness are modified at each step, so that it results that the generated structures are optimized in regard to different weights of the physical quantities.

2. Materials and methods

Genetic Algorithms [17,18] are a search method used to solve problems through three principal

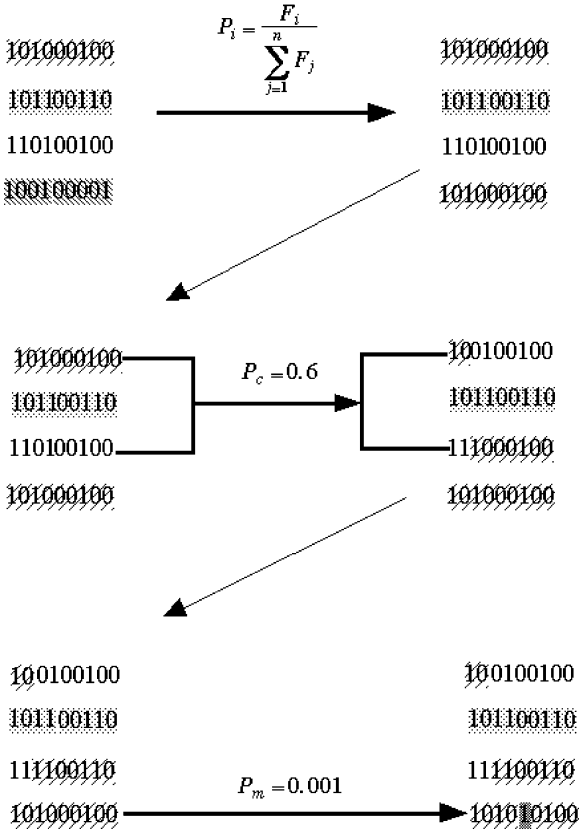


Fig. 1. Main steps of a Genetic Algorithm. From top to bottom: Reproduction, Crossing over and Mutation. The *P*'s are probabilities; the *F*'s are fitness values.

operators: selection, recombination and mutation of tentative solutions until the best one has been achieved (see Fig. 1). They represent a stochastic strategy able to handle the global minimum problem in a reasonable amount of time. The objective function which has to be optimized is allowed to have multiple local extrema, and to be high dimensional. These also are the characteristics of any protein potential landscape; therefore, the Genetic Algorithms seem to have a promising computational approach. Nevertheless, in the presence of a very high number of local minima even Genetic Algorithms stay trapped without reaching the optimum minimum, this phenomenon being called 'premature convergence'. Special operators exist to avoid or at least delay this event, and they will

be described later. But it can be more effective to operate upon the protein structure representation, as to obtain a less 'rugged' energy landscape. So, the operators of the Genetic Algorithm and the representation 'code' are linked aspects of the overall algorithmic tool.

The protein structure representation has been developed in the previous article [4], and we limit to recall its main features and properties here. Afterwards, we will completely describe the Genetic Algorithm and, finally, the searching procedure regarding the fitness function. The potentials used are briefly described in Appendix A.

2.1. The protein representation code

In the search space Genetic Algorithms sample more frequently subspaces characterized by a low mean value of the fitness function. This is the central idea of the so-called 'schemata theorem'. If a string of dihedral angle values is used to represent the protein structure, packed structures lie in regions of the search space with a high average values of the fitness function, which are due to the higher probability of clashes. This occurrence gives rise to the difficulty in reaching the minimum.

We adopt the discrete Haar transform of the backbone dihedral angles sequence to overcome this difficulty. Haar functions form a complete orthonormal function set of rectangular waveform [19] (see Fig. 2), and were originally proposed by Haar in 1910 [20]. When the structure is reconstructed, it is evident that the low degrees of the transform contain the information regarding the global character of the sequence (the secondary structure), whilst the high degrees contain the local distortions, which account for the global folding and the tertiary interactions. Then the Haar transform representation appears to be very different from the simple sequence of dihedral angles, since these accounts both for the local interactions and for the tertiary ones. Therefore, this representation should assist structural path evolutions, which mimics the transition through the molten globule ensemble, and should then result in a less frustrated energy landscape.

The immunoglobulin binding protein G (1IGD), a small (61 aminoacids) globular protein, has been chosen for the first attempts as it is the smallest of the database we used in the previous article [4]. Its high-resolution crystal structure was obtained from the Brookhaven protein data bank (PDB) [21]. In Fig. 3, a simplified view of the crystal structure and besides it, the reconstruction obtained by the inverse Haar transform using all the functions, except those with the highest degree, are shown. The formation of most of the secondary motifs in the reconstructed structure is evident: the dihedral angles fall in the proper α- or β-structure regions. On the other hand, it appears as an 'open' structure, and, therefore, it surely lacks the tertiary interactions.

We used the public domain program RasMol [22] for the visualization of molecules.

2.2. The genetic algorithm

Projecting a Genetic Algorithm designed to solve a particular problem involves specific operators (rules that modify the individuals and the population). Following in the next sections are the main definitions and operators that form the Genetic Algorithm specific for the problem of searching folded protein structures.

2.2.1. Starting population

The starting population has been generated by a random choice of ϕ and ψ dihedral values for each residue, weighted by the statistical distributions of the individual aminoacids extracted from the Brookhaven Protein Data Bank [23,24]. This is a knowledge-based approach and we do not use any secondary structure prediction. This procedure allows generating better initial conformations of proteins (that is, with a smaller amount of clashes) in contrast to a purely random choice.

Only the C_β atoms represent side chains.

Good results are obtained in the search using a population size ranging from 120 to 250 individuals.

2.2.2. Operators

The usual operators (reproduction, crossing over and mutation) have been used in their usual form,

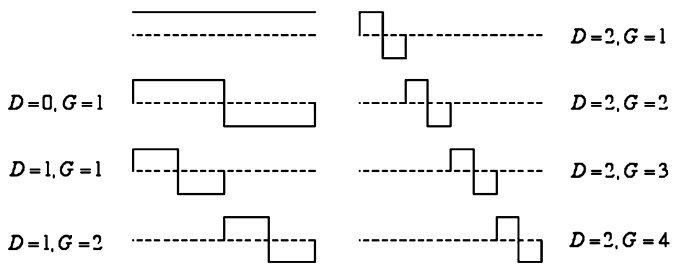


Fig. 2. The solid lines represent some mono-dimensional Haar functions. The function without symbols is Haar(0,t). The degree (D) and the order ($G=m+1$) of the functions Haar($2^D+m,t$) are specified. Functions with the same degree and different position have different order. These functions should be multiplied for$(\sqrt{2})^D$ to be reported on the same scale.

along with other versions adjusted to the peculiarity of the problem.

2.2.2.1. Sharing function. A sharing function has been introduced to keep the structures as different as possible. The 'distance' between two structures is defined as one minus the fraction of common residue conformations (for the assignment of a φ,ψ pair to the α or β conformation, see in Appendix A, the components of the fitness designed as 'Number of residues in α conformation' and 'Number of residues in β conformation'). The distance will be 0 for identical individuals and 1 for completely different ones. The sharing function s_i of i structure is the summation of its over all distances from all the individuals of the population. The fitness function is divided by the sharing function.

2.2.2.2. Scaling of the fitness function. Following a classical approach in Genetic Algorithms, the fitness function is scaled, in order to have a predetermined multiplicative reproduction coefficient C_m for the best member of the population [17]. We use the value $C_m=1.2$.

2.2.2.3. Crossing over (CO) operators. The CO operators are simultaneously applied to both the Haar coefficients for φ and ψ angles, in order to act on the overall residue conformation. Three

Fig. 3. 1IGD protein backbone represented in ribbon style. Right: native structure. Left: structure obtained by the inverse Haar transform, using all the functions except those with the highest degree. The α-helical fragments are black. The β-helical fragments are dark gray. The remaining residues are light gray.

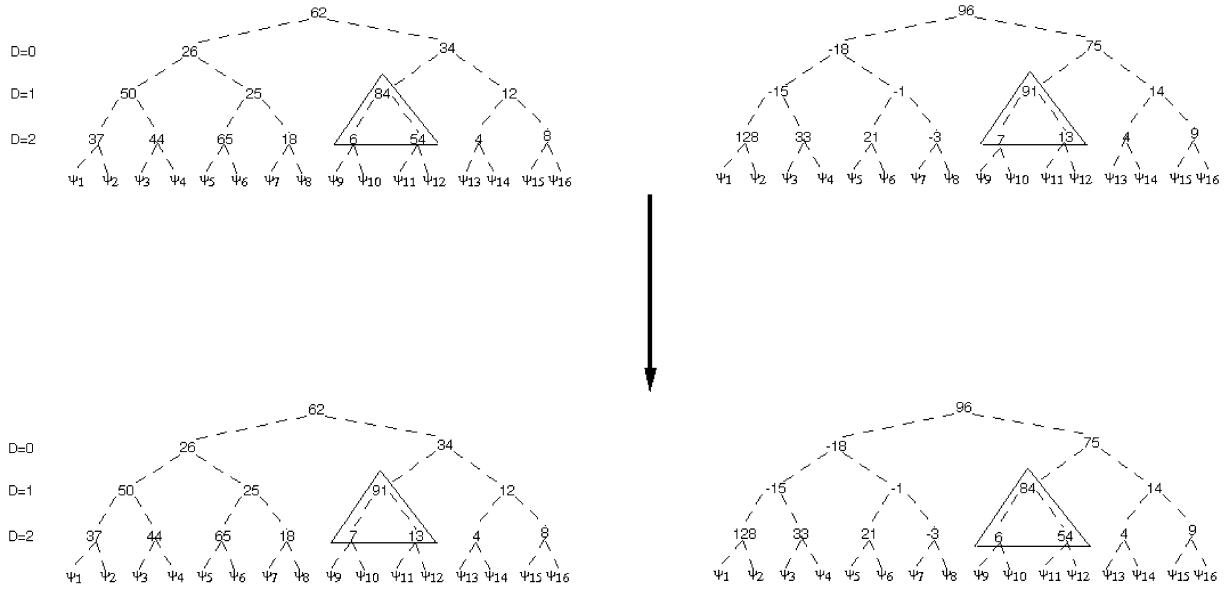


Fig. 4. 'Triangle' crossing over. Upper: an old generation pair. Down: a new generation pair. The triangles identify the Haar coefficients involved in the crossing over between the pair. The sketched lines ascending from the torsional angles ψ_i, link the Haar coefficients involved in their determination.

different kinds of CO operators are randomly chosen at each generation. Besides the one- and two-point crossover operators, which generally change all the conformations of the structure, we use a CO acting on a limited amount of torsional angles. This CO operator exchanges both a randomly chosen Haar coefficient and the ones in increasing order regarding the same residues ('triangle' crossing over) (see Fig. 4).

2.2.2.4. Mutation operators. Four different kinds of mutation operators are randomly chosen at each generation. Along with a random mutation operator, we have a mutation operator acting with a small increment to the Haar coefficient randomly chosen, in order to change the structure slowly. This last operator can be called 'variation' [25].

A mutation acting over a couple of Haar coefficients (where the ψ change is double the φ one) promotes a conversion among different local conformations, which is parallel to the line connecting α and β conformations in the Sheraga map.

A mutation attributing opposite random values to the Haar coefficients of a couple of φ, ψ angles will promote the β-turn formation and then the folding of the sequence.

2.2.2.5. New generation. We use the method called 'elitist generation replacement' [25]. The crossing over is performed to create a number of individuals equivalent to the size of the population. The parent individuals and the offsprings are then ranked by their fitness values. The better individuals are then chosen to form the new generation, excluding the duplicate ones.

2.2.2.6. Injection of new structures. If multiplicity is under a chosen value (i.e. 0.50), the injection of a new random structure will replace an old one randomly chosen at any iteration, so to avoid uniformity in the population. The multiplicity is obtained as the mean value of the sharing values, divided by the population size. This operator has been developed to reintroduce diversity into the population, as diversity is lost through the evolution.

2.2.3. Fitness function

Our interest does not reside as much on the effectiveness of the single potential functions, as on deriving a procedure to establish the relative weights of the functions used, in order to maximize the effectiveness of their linear combination and the possibility to drive the algorithm toward the native structure. We choose the functions in order to test the procedure in the hardest conditions, that is a large number of different physical quantities, some of which obtained as knowledge-based potentials, while the others are ad hoc criteria, having a more empirical and intuitive character (see for example the clash). In previous studies, weight values were found by many different simulation trials [10–12].

Several authors [10,11,26–31] make use of a simplified representation of proteins, in which the side chain has been omitted or represented by a single atom. We use C_β atom to represent the side chain. Obviously, some of the components of the fitness must account for the role played by the omitted side chains in stabilizing the protein structure. In our case, such components are the hydrophobicity (calculated following two different approaches), the hydration, the logarithm of the product of the probabilities assigned to the conformations of the residues and the total sum of the square of the z-scores of the backbone torsional angles, since different aminoacids refer to different clusters of angle distributions. The potentials used (some of which are original) are briefly described in Appendix A, and the resulting fitness function will be expressed as:

$$\begin{aligned}\text{fitness} = \{ & a_1 * \text{clash}C\alpha + a_2 * \text{clash}C\beta \\ & + a_3 * \text{hydrophobicitySDH} \\ & + a_4 * \text{hydrophobicity}HF + a_5 * \text{total}HB \\ & + a_6 * \beta HB \ + a_7 * \beta HB\text{distance} \\ & + a_8 * \text{density}C\alpha + a_9 * \text{fourthmoment}C\alpha \\ & + a_{10} * sfd + a_{11} * lfd + a_{12} * \text{zscore} \\ & + a_{13} * \log\pi\text{probabilities} \\ & + a_{14} * \text{solvationenergy} + a_{15} * \text{number}\alpha \\ & + a_{16} * \text{number}\beta\end{aligned}$$

where the quantities are reported in the same order as in Appendix A, and the a_i coefficients are to be fixed with the search procedure.

2.3. Search procedure

The procedure starts from an initial population of non-native structures (see Fig. 5a). A search direction is calculated to identify the native structure as an extreme point (see Fig. 5b). In order to identify uniquely the search direction, we have chosen the one that maximizes the distances of the plane crossing the native structure from the nearest points of the set. A search is performed by the Genetic Algorithm, which in general identifies another non-native structure with a better fitness function (see Fig. 5c). This structure is added to the initial set and the procedure iterated. Adding a new structure to the set usually modifies the search direction at each step (see Fig. 5d). The whole process can stop because of the two alternative outcomes: (1) it only retrieves the native structure; in this case, an effective fitness function is identified; (2) it proves the impossibility of its achievement by any linear combination of the proposed quantities. This last conclusion will be attained when the iterative process generates new structures surrounding on all sides in the hyperspace the native one, so that it is not an extreme point of the set.

The stop modality makes this procedure an effective tool in evaluating the possibility to uniquely identify the native structure by a given set of physical quantities.

2.3.1. Scaling and dimensionality of the searching space

The used quantities can have a large numerical variability among them, even more so when calculated for different proteins. In order to prevent the predominance of the numerically larger quantities over the others, we always make use of their z-scores, which is a scaling of the axes of the searching space. We calculate the mean and standard deviation values for the quantities from a starting population of 1000 individuals. This device also permits to perform a parallel search on

Fig. 5

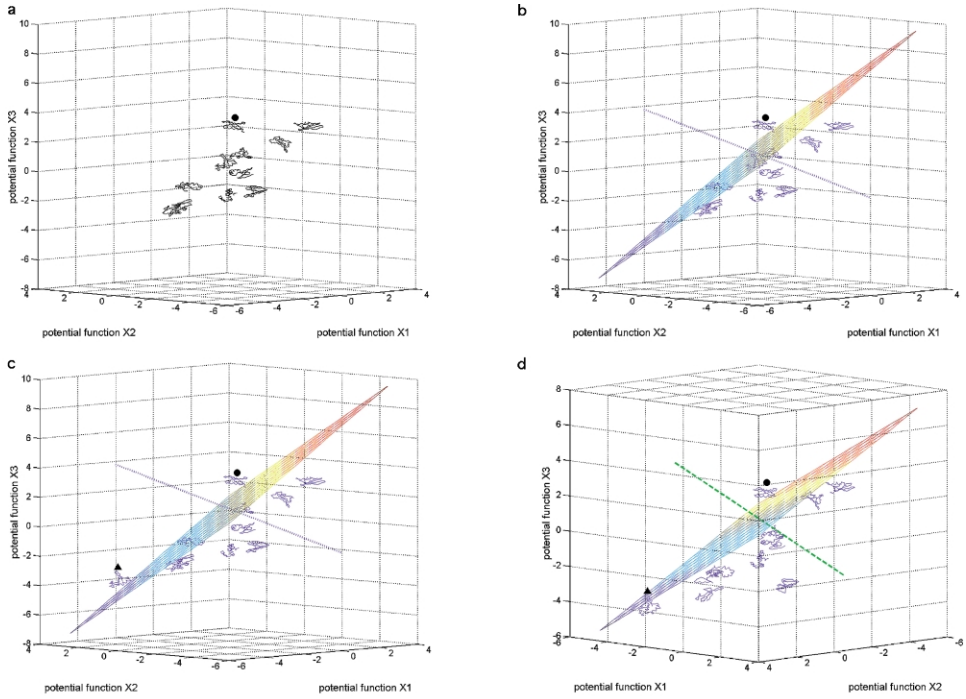
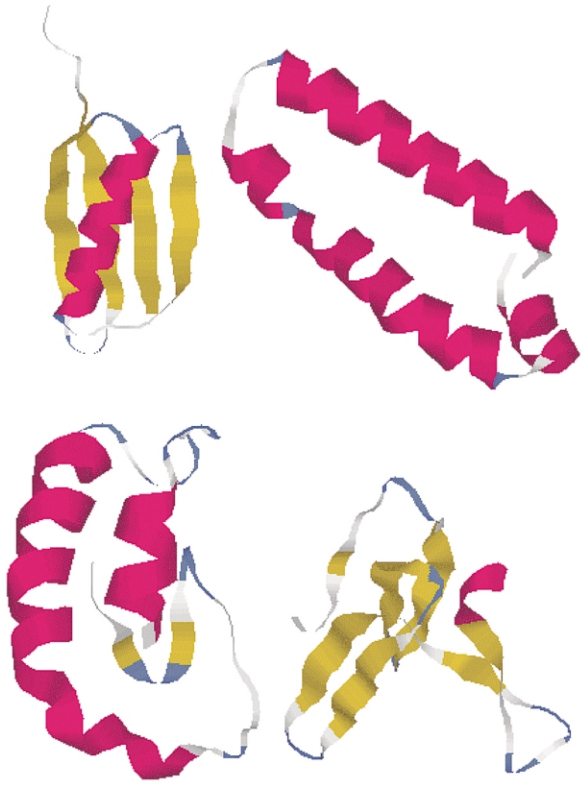

a
b
c
d
potential function X3
potential function X2
potential function X1

Fig. 6



more proteins, but we limit to the immunoglobulin binding protein G (1IGD) as for the first attempt.

It is in the interests of simplicity and of saving computational time to use the smallest number of variables. Besides, the quantities used can be correlated. Therefore, we begin the search for the fitness function with a reduced set of the quantities, and we raise their number when the search fails. The order followed in the choice is grounded over the Principal component analysis (PCA) [32]. We have obtained the correlation matrix of the numerical values of all the quantities considered from the server NetMul, a WWW online multivariate analysis system [32,33]. With the software Mathcad PLUS for Macintosh we have performed the principal component analysis, obtaining the eigenvectors, the eigenvalues and the component scores. The quantities used are ordered step-by-step following their descending component scores, as most of the capability to distinguish the structures is contained within the quantities with higher component scores.

2.3.2. Obtaining the searching direction

As long as the native structure is an extreme point of the set, a hyperplane including the native structure exists, that leaves all the other points representing the non-native structures on the same side. The line perpendicular to such a hyperplane has the property that the projection of the point representing the native structure over it is larger than the projections of all the other points. That is:

$$\sum_i a_i(x_i^s - x_i^j) > 0 \qquad (1)$$

where x_i is the value of the ith property, s identifies the native structure and j refers to a non-native set structure. The summation is extended to all the considered properties, and there will be a set of inequalities Eq. (1), one for each non-native structure, to be simultaneously fulfilled. The values a_i are both the line components and the weights of the quantities in the fitness function. The x_i are calculated values, whilst the a_i are the variables to be determined. Our choice of maximizing the projection is also a choice over the a_i signs. Changing the signs of all the a_i, merely switches from maximization to a minimization problem.

It is noteworthy that all the inequalities Eq. (1) contain a difference with the experimental structure, and that the x_i are the z-scores of the calculated quantities, so that the inequalities Eq. (1) can also include different proteins.

The set of inequalities Eq. (1) has in general a polyhedral cone of solutions [13]. We choose the direction that maximizes the projection differences with the nearest points, that is:

$$\sum_i a_i(x_i^s - x_i^j) > d \qquad (2)$$

where a high value of d is obtained by a few trial and error steps.

The set of inequalities Eq. (2) defines a linear programming problem. We have utilized NEOS server [34] of the Optimization Technology Center at Argonne National Laboratory and Northwestern University [35], designed to solve optimization problems remotely over the Internet. We have used PCx [36], which is an implementation of Sanjay Mehrotra's predictor-corrector interior-point linear programming package. The data have been given in the form of MPS files [37].

3. Results and conclusions

The immunoglobulin binding protein G (1IGD), a small (61 aminoacids) globular protein, has been

Fig. 5. The four steps of the iterative process in order to achieve a working fitness function are shown: (a) A set of misfolded structures together with the experimental one. The native structure is labelled with a point. (b) A hyperplane passing through the experimental structure and leaving the whole set of alternative ones on one side is shown. (c) Following the research direction (given by the perpendicular to the hyperplane), the Genetic Algorithm found a new structure (labelled with a triangle) better then the experimental one. (d) After solving the new inequalities system a new hyperplane, which leaves again the misfolded structure below it, and a new research direction are found.

Fig. 6. The 1IGD crystal structure (upper left corner) with some alternative structures having an equivalent fitness function. They are represented as ribbons. Red indicates α helix, yellow β sheet and blue β turn motif.

Table 1
Values of the final weights

Clash of C_α atoms	-1.0	Fourth square of the fourth moment of the C_α–C_α distance	0.27
Clash of C_β atoms	-1.0	Short range fractal dimension	-3.5×10^{-3}
Single residue hydrophobicity	3.5×10^{-2}	Long range fractal dimension	-5.7×10^{-3}
Hydrophobic fitness score	-9.6×10^{-3}	Total sum of the square of the z-scores of the backbone torsional angles	-0.20
Hydrogen bond energy	1.0	Logarithm of the product of the probabilities assigned to the conformations of the residues	-1.9×10^{-2}
Hydrogen bond in β sheets	-5.0×10^{-3}	Solvation free energy	0.10
Summation over all β aminoacids of the minimum deviation from the standard H bond distance	-0.79	Number of residues in α conformation	-0.58
Density of C_α atoms	6.0×10^{-2}	Number of residues in β conformation	-5.3×10^{-2}

chosen to test our procedure. A population size ranging from 120 to 250 individuals has been used. To control the Genetic Algorithm, we have previously applied it to the 1IGD protein, using the summation of the differences from the C_α distances and alternatively the differences from the dihedral angles of the native structure as fitness function. All the runs converged to a RMSD, which is less than 1 Å (calculated over the C_α coordinates), in less than 1000 iterations.

Given a fitness function, the purpose, in each step of the procedure, was to reach a structure with either an equal or a better fitness value than the one of the experimental structure. For all the fitness functions used, this aim has always been obtained with a number of iterations ranging from 1000 to 10 000, thus demonstrating the efficiency of the algorithm.

In order to test the procedure, we choose the potentials so to realize the hardest conditions, instead of making considerations over their validity as components of a working fitness function. Therefore, we use a large number (sixteen) of quantities, with a very different physical meaning, so to be substantially incommensurable.

After 310 iterations, there is no linear combination able to identify uniquely the native structure as the one with the best fitness value, notwithstanding the dimension of the set. Some structures with the same fitness value of the native one are shown beside the native structure in Fig. 6, and the values of the weights obtained in the last iterative step are shown in Table 1 (see Appendix A, for the mining of the quantities). Our opinion about the failure of the search is that the criteria related to the compactness of the structure are not very effective, as it can be deduced by the visual inspection of the structures obtained.

In conclusion, the described Genetic Algorithm and its specific operators set together with the used protein representation code result to overcome the problem of premature convergence. The topological point of view of the fitness function, together with a ‘dynamical’ generation of the decoys, comes out to be an effective tool in testing a given set of quantities, in order to establish if

they are able to be combined in a fitness function capable to identify the native structure of a globular protein.

Acknowledgments

The authors would like to thank Dr A. Scipioni for useful discussions. Grant sponsor: University of Rome 'La Sapienza'; Grant number:'Progetto 60% Facoltà'. Grant sponsor: MURST; Grant number: 'Cofinanziamento 40%'.

Appendix A: Components of the fitness function

A.1. Clash of C_α atoms

According to Dandekar and Argos [10], the clash criterion has been defined as: clash $= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Theta(i,j)$ where $\Theta(i,\ j)=1$ if $\sqrt{\sum_{k=1}^{3}(C_\alpha(res_i)_k - C_\alpha(res_j)_k)^2} \le 3.8$ Å else $\Theta(i,\ j)=0$; i, j are residue indexes, n is the total number of C_α in the protein under consideration, and k is the index of Cartesian coordinates.

A.2. Clash of C_β atoms

The same expressions used for C_α, but applied to C_β atoms.

A.3. Single residue hydrophobicity (SDH)

It is defined by Casari and Sippl [38,39]. The hydrophobic interaction energy $\Delta\Phi^{ab}(r)$ of the pair ab is: $\Delta\Phi^{ab}(r) \approx (h_a + h_b)(r - r_0)$ where r is the distance between the C_β atoms of amino acids a and b, and $r_0 = 23$ Å is a constant obtained from $\Delta\Phi^{ab}(r) = 0$. h_a and h_b are the hydrophobicities of the a and b residues, respectively.

A.4. Hydrophobic fitness score (HF)

It is calculated following the definition of Huang et al. [40,41]: $-\dfrac{(\sum_i B_i)[\sum_i (H_i - H_i^0)]}{n^2}$ where the summation is over all hydrophobic (C, F, I, L, M, V, W) residues i; B_i is the number of side-chain centroids within 10 Å; H_i is the number of non-polar centroids within 7.3 Å; H_i^0 is the number of hydrophobic contacts expected on a random basis; n is the number of hydrophobic residues in the sequence.

A.5. Hydrogen bond energy

The closest nitrogen atom to each main-chain carboxyl oxygen has been taken. We use the Kabsch and Sander's [42] expressions to evaluate the interaction energy:

$$E = q_1 q_2 \left(\frac{1}{r(\mathrm{ON})} + \frac{1}{r(\mathrm{CH})} - \frac{1}{r(\mathrm{OH})} - \frac{1}{r(\mathrm{CN})} \right) f$$

where q_1 and q_2 are partial charges on the C, O $(+q_1,\ -q_1)$ and N,H $(+q_2,\ -q_2)$ atoms, with $q_1 = 0.42$ e, $q_2 = 0.20$ e, e being the unit electron charge, and r(AB) the interatomic distance from A to B. r is in Å, E is in Kcal/mol, and the dimensional factor $f = 332$.

A.6. Hydrogen bond in β sheets

The interaction energy of hydrogen bonds in β sheets is evaluated applying the formulation of Kabsh and Sander to pairs of residues in β conformations. For the assignment of a φ,ψ pair to the β conformation, see below: 'Number of residues in β conformation'.

A.7. Summation over all β aminoacids of the minimum deviation from the standard H bond distance

The modulus of the difference between the distances and the standard value of 1.8 Å is performed. The minimum value of the modulus is taken for each residue and the summation is extended over all the β residues.

A.8. Density of C_α atoms

We have previously studied the density of the first momentum of C_α distribution [43]. Since the distribution of atoms in proteins is generally anisotropic, averages over spherical shells are inappropriate. Therefore, we make a metric deformation, which leaves the density invariant, so that the distribution is formally isotropic. Reduced vectors are introduced, defined as:

$$s = (\langle r_1^2\rangle\langle r_2^2\rangle\langle r_3^2\rangle)^{\frac{1}{6}} \left(\frac{r_1}{\langle r_1^2\rangle^{\frac{1}{2}}} \frac{r_2}{\langle r_2^2\rangle^{\frac{1}{2}}} \frac{r_3}{\langle r_3^2\rangle^{\frac{1}{2}}} \right)$$

where r_1^2, r_2^2, r_3^2 are the principal axes of the tensor of the second moment of the vector r of C_α atoms. The tensor and its diagonal form are, respectively:

$$\begin{vmatrix} \langle r_x^2 \rangle & \langle r_x r_y \rangle & \langle r_x r_z \rangle \\ \langle r_y r_x \rangle & \langle r_y^2 \rangle & \langle r_y r_z \rangle \\ \langle r_z r_x \rangle & \langle r_z r_y \rangle & \langle r_z^2 \rangle \end{vmatrix} \qquad \begin{vmatrix} \langle r_1^2 \rangle & 0 & 0 \\ 0 & \langle r_2^2 \rangle & 0 \\ 0 & 0 & \langle r_3^2 \rangle \end{vmatrix}$$

Using s vectors, the radial density has the constant value of 6×10^{-3} $C_\alpha/\text{Å}^3$ over the 90% of the radial coordinate and drops near the enveloping surface. The difference between the calculated and the expected density values is used in the fitness function.

A.9. Fourth square of the fourth moment of the C_α–C_α distance

It is defined as: $[\frac{2}{n}(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} r_{ij}^4)]^{\frac{1}{4}}$ where r_{ij} is the distance between the C_α atoms of amino acids i and j, and n is the number of residues of the protein.

This quantity is dimensionally a length. The higher is the exponent of the positive moment of the C_α–C_α distances, the more sensitive it is to high values. Consequently, lowering moments' values has the effect to 'drive' the sequence toward compact structures, which can then be used as another criterion acting similarly to the density. Calculating the rising moments (actually the ith square of the ith moment), a tendency toward an asymptote is shown for a set of native crystal structures, so that we choose the fourth moment as a compromise between a high sensitivity to the high values of C_α–C_α distances and the flattening on the asymptote.

A.10. Short range fractal dimension (sfd) and long range fractal dimension (lfd)

The length of the protein backbones has been studied as the function of the scale, where the length unit has been defined as the number m of consecutive C_α [44]. The obtained fractal diagram consists of two components, one for $m \leq 10$, characterized by a fractal dimension $sfd = 1.3397$ with a standard deviation of 0.0517, and the other for $m > 10$, characterized by a fractal dimension $lfd = 1.9536$ with a standard deviation of 0.1725. The first (sfd) should reflect the local (secondary) folding of the protein, whereas the latter (lfd) should reflect the global folding. The components of the fitness are given by $|sfd - 1.34|$ and $|lfd - 1.95|$, respectively.

A.11. Total sum of the square of the z-scores of the backbone torsional angles

The experimental PDB values of the torsional angles φ and ψ reported on the Ramachandran map, well represent the differences among the residues in the preference of particular zones of the map. A more recent statistical analysis was carried out by S.L. Sclove [23,24], who fitted the experimental data by a Bayesian finite mixtures of Gaussian clusters. If we assume that a cluster of experimental data also reflects the Boltzmann distribution, it follows that ΔE (the energy difference between a particular conformation φ_i, ψ_i from the main values of the cluster φ_j, ψ_j) is proportional to the 'distance' from the conformation represented by the main values, expressed in z-score terms:

$$\Delta\Theta_i = \left(\frac{\psi_i - \psi_j}{\sigma_{\psi_j}}\right)^2 + \left(\frac{\varphi_i - \varphi_j}{\sigma_{\varphi_j}}\right)^2$$

where σ_{ψ_j} and σ_{φ_j} are the standard deviations of the cluster under consideration.

The φ_i, ψ_i conformation is attributed to the cluster from which it has the minimal distance in standard deviation units. A summation of terms $\Delta\Theta_i$ is performed over all the aminoacids i of the sequence.

A.12. Logarithm of the product of the probabilities assigned to the conformations of the residues

The directional information measure introduced by Garnier, Osguthorpe and Robson [45], is used to calculate the probabilities that each single residue in the sequence assumes the α, β, β-turn or coil conformation. In the structure under consideration each residue will be assigned to a conformation, either following the assignment procedure described in the fitness components, which are

designed as 'Number of residues in β conformation' and 'Number of residues in α conformation', or following the classification of Chou and Fasman [46] along with the analysis of the adjacent residues for the β-turn assignment. The product of the probabilities regarding the residue conformations is performed, and its logarithm is calculated, to extract a quantity of the same kind of the previous secondary structure propensity measure, and following reasoning similar to the previous component.

A.13. Solvation free energy

The solvation free energy has been estimated following the criteria of Huang et al. [40,41], and the experimental ΔG° values of hydration [47]. Simple geometrical considerations indicate that a maximum number of three C_β atoms can surround the C_β under consideration in a tetrahedral arrangement if the C_α of the residue is also considered. If H_i^β is the number of C_β within 7.3 Å from the C_β of the residue i, then $(3-H_i^\beta)/3$ is a rough estimation of the percentage of surface exposed to the solvent, and the solvation free energy for the structure considered can be expressed as: $\sum_i \frac{(3-H_i^\beta)}{3}\Delta G_k^0$ where the summation is the over all residues, and k indicates what kind of amino-acid the residue i is.

A.14. Number of residues in α conformation

We have referred to the main values and standard deviations for the torsional angles such as those extracted by the statistical mixture model analysis [24] for the α and β structures, using the reported frequencies and the well-known relationships among the moments [43]. $A \pm 2\sigma$ dispersion around the main value has been admitted. Therefore, the admitted intervals for the φ, ψ angles for residues in a regular right-handed α helix are $(-78^\circ, -50^\circ)$ and $(-54^\circ, -22^\circ)$, respectively. The α conformation is included in the summation only if it is merged in at least four consecutive α conformation residues, to allow the H-bond formation.

A.15. Number of residues in β conformation

The admitted intervals for the φ, ψ angles for residues in a regular β-sheet helix are $(-157^\circ, -89^\circ)$ and $(110^\circ, 166^\circ)$, respectively. Following Kabsh and Sender [42], isolated β conformations are not taken into account.

References

[1] J. Bryngelson, J.O. Onuchic, N.D. Socci, P.G. Wolynes, Funnels, pathways and the energy landscape of protein folding: a synthesis, Proteins Struct. Funct. Genet. 21 (1995) 167–195.

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

[3] J. Bryngelson, P.G. Wolynes, Spin glasses and the statistical mechanics of protein folding, Proc. Natl. Acad. Sci. USA 84 (1987) 7524–7528.

[4] S. Morosetti, Discrete Haar transform and protein structure, J. Biomol. Struct. Dyn. 15 (1997) 489–497.

[5] R. Jaenicke, Stability and folding of domain proteins, Prog. Biophys. Mol. Biol. 71 (1999) 155–241.

[6] J.N. Onuchic, Z. Luthey-Schulten, P.G. Wolynes, Theory of protein folding: the energy landscape perspective, Annu. Rev. Phys. Chem. 48 (1997) 545–600.

[7] V.P. Denisov, B.H. Jonsson, B. Halle, Hydration of denatured and molten globule proteins, Nat. Struct. Biol. 6 (1999) 253–260.

[8] A.A. Rabow, H.A. Scheraga, Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator, Protein Sci. 5 (1996) 1800–1815.

[9] Y. Cui, R.S. Chen, W.H. Wong, Protein folding simulation with genetic algorithm and supersecondary structure constraints, Proteins Struct. Funct. Genet. 31 (1998) 247–257.

[10] T. Dandekar, P. Argos, Folding the main chain of small proteins with the genetic algorithm, J. Mol. Biol. 236 (1994) 844–861.

[11] T. Dandekar, P. Argos, Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using genetic algorithm and extended criteria specific for strand regions, J. Mol. Biol. 256 (1996) 645–660.

[12] T. Dandekar, P. Argos, Applying experimental data to protein fold prediction with the genetic algorithm, Protein. Eng. 10 (1997) 877–893.

[13] K.S. Kölbig, F. Schwarz, A program for solving system of homogeneus linear inequalities, Comput. Phys. Commun. 17 (1979) 375–382.

[14] S. Lang, Linear Algebra, Addison-Wesley, Reading, MA, 1966.

[15] B. Fain, Y. Xia, M. Levitt, Design of Chebyshev-expanded discrimination function for protein structure prediction, Protein Sci. 11 (2002) 2010–2021.

[16] M. Vendruscolo, R. Najmanovich, E. Domany, Can a pairwise contact potential stabilize native protein folds

against decoys obtained by threading?, Proteins Struct. Funct. Genet. 38 (2000) 134–148.
[17] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[18] S. Forrest, Genetic algorithms: principles of natural selection applied to computation, Science 261 (1993) 872–878.
[19] K.G. Beauchamp, Applications of Walsh and Related Functions, Academic Press, London, 1984.
[20] A. Haar, Zür theorie der orthogonalen funktionensisteme, Math. Annal. 69 (1910) 331–371.
[21] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, et al., The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[22] RasMol, Version 2.6. Edinburgh: Biocomputing Research Unit; Glaxo Research and Development, Greenford, UK, 1993.
[23] S.L. Sclove, S.A. Sherman, Cluster Analysis of Dihedral Angles of Amino Acid Residues in Proteins. in: 1994 Proceedings of the Biopharmaceutical Section (Ed.), American Statistical Association (American Statistical Association, Alexandria, VA, 1995), pp. 399–404.
[24] S.L. Sclove, S.A. Sherman, A priori and a posteriori mixture distributions for using databases in protein structure determination, J. Mol. Struct. (Theochem.) 419 (1997) 245–255.
[25] S. Schulze-Kremer, Molecular Bioinformatics. Algorithms and Applications, Walter de Gruyter, Berlin, 1996.
[26] S. Sun, Reduced representation model of protein structure prediction: statistical potential and genetic algorithms, Protein Sci. 2 (1993) 762–785.
[27] A. Wallqvist, M. Ullner, A simplified amino acid potential for use in structure predictions of proteins, Proteins 18 (1999) 267–280.
[28] F. Herrmann, S. Suhai, Genetic algorithm in protein structure prediction, in: S. Suhai (Ed.), Computational Methods in Genome Research, Plenum Press, New York, 1994, pp. 173–190.
[29] R. Unger, J. Moult, Genetic algorithms for protein folding simulations, J. Mol. Biol. 231 (1993) 75–81.
[30] J.U. Bowie, D. Eisenberg, An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function, Proc. Natl. Acad. Sci. USA 91 (1994) 4436–4440.
[31] F. Herrmann, S. Suhai, Energy minimization of peptide analogues using genetic algorithms, J. Comp. Chem. 16 (1995) 1434–1444.
[32] HTTP://pbil.univ-lyon1.fr/ADE-4/NetMul.html.
[33] J. Thioulouse, F. Chevenet, NetMul, world-wide web users interface for multivariate analysis software, Comput. Stat. Data Anal. 21 (1996) 369–372.
[34] http://www-neos.mcs.anl.gov/.
[35] http://www.ece.nwu.edu/OTC/.
[36] http://www-fp.mcs.anl.gov/otc/Tools/PCx/.
[37] http://www-fp.mcs.anl.gov/otc/Guide/OptWeb/continuous/constrained/linearprog/mps.html.
[38] G. Casari, M.J. Sippl, Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds, J. Mol. Biol. 224 (1992) 725–732.
[39] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, 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 (1990) 167–180.
[40] E.S. Huang, S. Subbiah, M. Levitt, Recognizing native folds by the arrangement of hydrophobic and polar residues, J. Mol. Biol. 252 (1995) 709–720.
[41] E.S. Huang, S. Subbiah, M. Levitt, Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations, J. Mol. Biol. 257 (1996) 716–725.
[42] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[43] P. De Santis, S. Morosetti, A. Palleschi, Moments of the distribution of the amino acid residues in tertiary structures of globular proteins, Biopolymers 18 (1979) 2963–2978.
[44] I. Yoshinori, I. Toshiyuki, Fractal analysis of tertiary structure of protein molecule, J. Phys. Soc. Jpn. 53 (1984) 2162–2171.
[45] J. Garnier, D.J. Osguthorpe, B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, J. Mol. Biol. 120 (1978) 97–120.
[46] P.Y. Chou, G.D. Fasman, Prediction of the secondary structure of proteins from their amino acid sequence, Adv. Enzymol. 47 (1978) 45–148.
[47] K.E. VanHolde, W.C. Johnson, P.S. Ho, Principles of Physical Biochemistry, Prentice Hall, Upper Saddle River, New Jersey, 1998.