Scaling laws for folding native protein structures

Fan Jiang*

Beijing National Laboratory of Condensed Matters, Institute of Physics, Chinese Academy of Sciences, Beijing 100080, China

Received 28 February 2005; accepted 30 May 2005
Available online 12 October 2005

Abstract

We propose a nucleation hypothesis for protein folding. Based on this hypothesis, we have designed a new nearest-neighbor method for the prediction of protein secondary structures, in which the reliability of each prediction is estimated based on sequence conservation and clustering in the databases of known structures. We have found that predictions with higher reliability scores were indeed correlated with a higher predictive accuracy. We also found that by selecting the top 20% of residues based on reliability scores as nucleation residues, a clear pattern emerged where hydrophobic amino acids were largely buried and hydrophilic amino acids were more exposed. This was consistent with the widely accepted HP-model for protein folding. These results were true for several databases, such as PDBSELECT (<25% sequence homology), SCOP-ASTRAL (<25% sequence homology), and SCOP-ASTRAL unique fold classes, with 1300, 3956, and 762 proteins, respectively. Therefore, it is conceivable that the nucleation residues function not only as initiation sites for folding, but also as the core residues playing a primary role in determining the protein (thermodynamic) stability. The occurrence of these two functions on one set of amino acid residues in a protein is perhaps from the result of biological evolution. Finally, we have found power law behaviors in our results, whose scaling properties were modeled using polymer physics and critical phenomena. It is concluded that proteins behave like a real chain. A new physical picture for protein folding derived from our nucleation hypothesis has been described, in which there are two continuous phase transitions corresponding to the two stages of protein folding: one is nucleation and the other collapse. By determining the critical exponents of these two-phase transitions, it has been found that the nucleation process has a spatial dimension of $d = 3$ while the collapse process of $d = 2$.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Protein folding; Protein thermodynamic stability; Protein kinetics; Secondary structure prediction; Nucleation; Collapse; Finite size phase transition; Critical exponent; Scaling property; Polymer physics

1. Introduction

The central dogma of molecular biology has been the focus of biological studies for the last three decades. The goal is to understand the nature of genetic information contained in a cell. It is now known that the main function of this genetic information is to encode protein sequences and their associated expression networks [1,2]. These molecular components assembled together make a cell function as it does. It has been found that the structure and function of a protein is determined by its amino acid sequence, in particular, by Anfinsen [3], in the in vitro unfolding and refolding of ribonuclease A [1,3]. The second part of the central dogma, or the second half of the genetic code, refers to the question of how the amino acid sequence of a protein determines its structure [4]. This question was described as a paradox by Levinthal [5]. He analyzed the kinetic problem which a protein encounters when folding and pointed out that the necessity to search through a huge number of conformations of a polymer chain, such as a protein, in order to reach its thermodynamic equilibrium state as required by Anfinsen’s principle, would result in the folding process taking a prohibitively long time ($10^{36}$ yrs). In this paper, we will attempt to show that proteins fold in general through a step-wise process.

Our working hypothesis was the following: there exists so-called nucleation residues and similar nucleation residues exist in all globular proteins. What we mean here by ‘similar’ is that the amino acid sequences of these nucleation residues are more conserved than the residues across all other sequences. Based on this hypothesis, the folding process is described as the following: the folding
initiates with the formation of some specific secondary structures, whose formation is controlled by short-range interactions. These pre-formed secondary structures will lower the effective entropy barrier for folding and limit the number of possible conformations and/or folding pathways to be sampled in the subsequent folding. Then, the folding proceeds to the formation of secondary and tertiary structures whose formations are dictated by tertiary or long-range interactions. Long-range interactions here refer to interactions between two residues separated far apart in a sequence. These tertiary interactions will further stabilize the secondary structures formed at the initiation stage. It should be noted that the pre-formation of specific secondary structures is in the sense that there is a strong tendency or probability for its formation. More importantly, this initial tendency is preserved towards the end of the folding process so that it results in the same specific secondary structures as the final native structure. This immediately offers a simple way to test our working hypothesis. That is a method needs to be designed to find the so-called nucleation residues whose secondary structures could be predicted with a much higher accuracy than the average for all residues in a protein.

Since the nucleation residues as proposed are related to only short-range interactions, only continuous segments of sequences need to be examined. These short segments could be called ‘foldons’. Therefore, a nearest-neighbor type of algorithm was designed, which has been described previously [6]. In the previous work, it was found that the predictive accuracy of secondary structures correlated well with the reliability z-score. Here, two more databases have been tested and similar results were obtained. The solvent accessible surface area analysis showed that for the top 20% most reliable predictions, the relation between the average solvent accessible surface areas and the hydrophobicities for the 20 amino acids was particularly interesting. The conclusion was consistent with the HP-model for protein folding in lattice systems [7]. Further statistical analysis led to the discovery of power laws in the distribution of residues with respect to the reliability z-score. Based on the scaling properties of a real chain in polymer physics as described by de Gennes [8], it was found that the critical exponents of the power laws could be determined in terms of spatial dimensions only. These findings led to a new view for the protein folding process, which seemed to be consistent with the overall physical picture derived from the nucleation hypothesis.

2. Materials and methods

2.1. Protein structure databases

Two new databases were used in the present work and they were derived from SCOP-ASTRAL version 1.63 (URL: astral.berkeley.edu) [9,10]. One database contained unique fold classes with 762 proteins and the other had all proteins with mutual sequence homology less than 25%, which had 3965 entries after deleting entries with missing sequences or containing only CA atoms. Chain breaks were allowed, although the 762 entries for unique fold classes had no chain breaks. Both X-ray and NMR structures were used. For NMR structures, the average model, if provided, or the first model in an entry was used. For all the figures and the results presented in this work, the data was derived from the database of 3965 entries unless otherwise specified.

2.2. Definition of the secondary structures and the three-state classification

The secondary structures of each entry, whose main chain atom coordinates must be known, was determined with the DSSP program [11], which uses an eight-state assignment for secondary structures. This was converted to the three-state assignment by a commonly used convention [12], containing helix, sheet, and coil.

2.3. Incorporation of BLAST PROFILE in the secondary structure prediction

The current version of the BLAST search method has implemented the PSI-BLAST profile method [13]. The default parameters were used for generating the profile for each sequence in our database. The known sequence databases were downloaded from EBI (URL: ftp.ebi.ac.uk) and UniProt Release 1.4 (consisted of Swiss-Prot Release 42.11 of 01-Mar-2004 and TrEMBL Release 25.11 of 01-Mar-2004).

2.4. Secondary structure prediction algorithm and the reliability score

The secondary structure prediction algorithm has been described previously [6]. It is briefly described in the following. For each protein in the SCOP-ASTRAL database, its sequence is divided into overlapping segments of equal length, say 17 residues. Since the secondary structure of each residue is known, each segment could be sorted into three different files according to the secondary structure state of the center residue of the segment, using the three-state secondary structure assignment (helix, sheet, and coil) mentioned above. So depending on the database used, the number of sequence segments contained in each of the three files was different. For example, for the SCOP-ASTRAL database with sequence homology less than 25%, containing 3965 proteins, the number of sequence segments for files helix, sheet, and coil was 254,896, 152,082, and 295,363, respectively. For a query sequence segment, a similarity z-score $z_i$ is defined as

$$z_i = \frac{\mu_{\text{seq},i} - \mu_{\text{random},i}}{\sigma_{\text{random},i}}$$  

(1)
where $\mu_{\text{seq},i}$ is the average similarity score of the query sequence segment to the database file ID, ID = 0, 1, 2 for helix, sheet, and coil, respectively. $\mu_{\text{seq},i}$ is calculated in the following way. The similarity score, between the query sequence segment and any known sequence segment in file ID, is calculated using the similarity matrix BLOSUM62 [13], a 20 by 20 matrix. Note that in this calculation no gaps are allowed. For example, for ID = 0, there are 254,896 scores, the same as the number of known sequence segments in file ID = 0. Then all these similarity scores for file ID = 0 are sorted in the descending order and the top 80 scores are averaged to give $\mu_{\text{seq},i}$. $\mu_{\text{random},i}$ is calculated in the same way except the query sequence segment is replaced by a random sequence segment. Usually more than 300 random sequence segments were generated and both the average $\mu_{\text{random},i}$ and the standard deviation $\sigma_{\text{random},i}$ were calculated using the top 80 similarity scores. After $z_i$, $i = 0, 1, 2$, are calculated, or let them be called $z_{\text{helix}}, z_{\text{sheet}}, z_{\text{coil}}$, the predicted secondary structure is the one with the highest $z$-score. The reliability score for this prediction is defined as:

$$z_{\text{reliability}} = \max\{z_i, i = 0, 1, 2\} - \max_{2nd}\{z_i, i = 0, 1, 2\}$$  \hspace{1cm} (2)

The accuracy of prediction is assessed by $Q_3$

$$Q_3 = \frac{\sum_{i=0,1,2} \text{Number of correct prediction for state } i}{\text{Total number of observation}}$$  \hspace{1cm} (3)

The accuracy for each individual state $i$, $i$ = helix, sheet, or coil, is $Q_i$

$$Q_i = \frac{\text{Number of correct prediction for state } i}{\text{Number of observation for state } i}$$  \hspace{1cm} (4)

2.5. Calculation of the solvent accessible surface areas

The solvent accessible surface area for each residue was extracted from the DSSP output [11]. In the statistical analysis presented in Section 3, the N-terminal and C-terminal four residues were excluded, as they were often less reliably determined in the experiments due to the more flexible conformations found either in X-ray or NMR structures.

3. Results and discussion

3.1. Correlation between the prediction accuracy and the reliability score

All predictions in this work have utilized the BLAST profile method [13]. The prediction accuracy in $Q_i, i$ = helix, sheet, coil is plotted against the reliability score $z_{\text{reliability}}$ in Fig. 1. It is clearly evident that both $Q_{\text{helix}}$ and $Q_{\text{sheet}}$ reached their maximum of 100% quickly as $z_{\text{reliability}}$ increases. Although $Q_{\text{coil}}$ increased continuously also, it did not reach 100%. Fig. 1 demonstrates that the reliability score as defined correlated very well with the prediction accuracy.

3.2. Statistical analysis of the solvent accessible surface areas

The average prediction accuracy $Q_3$ for all the sequences from both the database of unique fold classes and that of sequence homology <25% was 69.8 and 73.2%, respectively. In contrast, when the top 20% most reliable predictions for each protein sequence were examined, the $Q_3$ increased to 90.5 and 92.7%, respectively. So clearly the higher the reliability score of a prediction, the higher the prediction accuracy. More interestingly, when the solvent accessible surface areas were examined for the residues from the top 20% most reliable predictions, a clear trend in relation to the hydrophobicity of the 20 amino acids was observed. This is shown in Fig. 2, where the relative solvent accessible surface area for each amino acid is plotted in four

![Fig. 1. Relationship between the prediction accuracy and the reliability z-score.](image1)

![Fig. 2. Trend of the relative solvent accessible surface area as a function of the hydrophobicity of the 20 amino acids.](image2)
bins of reliability scores corresponding to the top 0–10, 10–20, 20–30, and 30–40% of the most reliable predictions. The reference solvent accessible surface areas for the 20 amino acids were calculated based on a set of accurate crystal structures published previously [14]. Hydrophobic amino acids are approximately defined as LIVAWFYMCTS and hydrophilic amino acids as GPHDENQRK, similar to that by Dill [7]. From Fig. 2, it can be seen that for the top two bins (0–10 and 10–20%), the trend was that hydrophobic residues were more buried than the average and, in contrast, hydrophilic residues were more exposed than the average. For the next two bins (20–30 and 30–40%), the trend was still valid but with some exceptions. This was also true for the data from the SCOP-ASTRAL unique fold classes, although with more fluctuations for the second two bins. Therefore, for the top two bins, the trend was the same for both databases. These results were the same as the previous result [6] using a database from PDBSELECT with sequence homology <25%.

The above results were surprising but within our expectation and seemed to be consistent with the notion that protein folding is driven by hydrophobic interactions [1,2,7]. It may be concluded that the nucleation residues as hypothesized were indeed found by our prediction method, as supported by the corresponding remarkably high prediction accuracy. These nucleation residues correspond, most probably, to the top 20% most reliably predicted residues. More importantly, it has been found that these nucleation residues exhibited a clear and consistent pattern of hydrophobic inside and hydrophilic outside, which would not have been seen if the whole sequence (100% of the residues in each protein) had been examined. As we know, more hydrophobic inside and more hydrophilic outside would lead to a more stable protein. Therefore, the nucleation residues as proposed here determine not only the protein folding kinetics but also the protein thermodynamic stability. It is remarkable that both the kinetic and thermodynamic requirements of a functional protein are satisfied with one and the same set of amino acid residues in its sequence. In other words, proteins through evolution have been able to unify both the physicochemical principles (kinetic) and the biological principles (thermodynamic). Coincidently, the current finding seems also to conform to the well-known 20–80 rule.

3.3. Power law behavior

Is there a power law behavior in protein folding? Using the above data, the logarithm for the number of residues \( N(z) \) as a function of the reliability score is plotted as a histogram with a bin size of 0.2 for the reliability score. In Fig. 3, the distribution of helix residues as a function of \( z \)-score can be modeled as two lines and the two lines join at a \( z \)-score approximately equal to 1.5. The distribution of sheet residues displays the same pattern. But the distribution of coil residues is rather different from helix and sheet, showing just one straight line. This is understandable because helix and sheet are secondary structures with regular hydrogen bonding patterns and their structures are more rigid than coil. Their formation depends more on the specific patterns in the amino acid sequence, whereas coils are in general more flexible and have less specific patterns in the amino acid sequence. This is consistent with Fig. 1, where the predictive accuracy for both helix and sheet can reach 100% while coil cannot. The distribution of all residues (sum of residues in all three states: helix, sheet, and coil) should more or less preserve the pattern in helix and sheet. By fitting data points 3–8 to one line and 10–15 to another, with data point 1 having a \( z \)-score of zero, the slope is found to be \(-0.533\) and \(-0.836\) with the corresponding \( R=0.996 \) and 0.999, respectively.

To demonstrate the fact that there is more than one physical process in effect dictating the formation of the secondary structures, especially helix and sheet, the distribution of the number of residues, whose secondary structures are determined by the tertiary interactions, is plotted in Fig. 4, again as a semi-log histogram plot. The distinction between residues whose secondary structures are determined by either short-range or long-range (tertiary) interactions is possible with our prediction method. Those that are predicted correctly using only information from the local flanking sequences of segments are considered the former. Those that are not predicted correctly are considered the latter. Surprisingly, all three states of secondary structures, helix, sheet, and coil, display a clear pattern of power law. This is consistent with the expectation that the formation of tertiary structures in protein folding is similar to a phase transition, as observed for example in differential scanning calorimetry (DSC) [15].

A new physical picture for protein folding can be derived from the above results. The folding starts with the formation of some specific secondary structures, mainly helix and/or sheet, determined by the nucleation residues...
as encoded in the sequence of a protein. Then these pre-formed secondary structures will collide and condense together to trigger the second stage of folding in which more secondary structures as well as tertiary structures start to form. Once triggered, the second stage of protein folding is characterized by collapse and annealing, where the long-range interactions take effect and dominate the formation of both the secondary and tertiary structures. The subsequent annealing and exchange of microscopic conformations to achieve close packing within the protein interior and to reach the thermodynamic minimum is the last and also the slowest step in protein folding. This physical picture is consistent with most of the experimental results published so far (see [16] for reviews). In this depiction of the physical process of protein folding, the transition state and the molten globule state, although they are not the same, should occur somewhere in between the nucleation stage and the collapse stage, either near the end of the nucleation stage or near the beginning of the collapse stage. It should be emphasized that the role of the nucleation stage is to reduce the effective entropy barrier for folding. Once the entropy barrier is overcome, folding is essentially downhill, so the collapse here is most likely a downhill hydrophobic collapse. From the results of the statistical analysis of large databases, it seems that nucleation is most likely the most common mechanism in reducing the entropy barrier. However, when a protein is relatively small and its chain topology is not very complicated, the effect of nucleation can vary and become very small because the entropy barrier for folding is not that large to start with. This might explain why many experiments do not indicate the importance of nucleation, especially for proteins which fold via a two-state process.

3.4. Estimation of critical exponents

It is clear that the collapse stage of protein folding could be considered a second order phase transition as evidenced by a single power law in the semi-log plot in Fig. 5. But the nucleation stage is usually considered a first order phase transition. Since a protein is not an infinite system, the effect of the finite size of a protein has been observed in the differential scanning calorimetry experiment [15], where the width of the delta function is broadened. Therefore, the nucleation stage might be considered a first-order continuous phase transition [17]. For simplicity, both the nucleation and collapse stages are modeled as second order phase transitions which exhibit critical phenomena.

de Gennes in Chapter X of his book ‘Scaling concepts in polymer physics’ [8] has demonstrated nicely the correspondence between polymer statistics and critical phenomena in general. We will use the relevant results here. First, the reliability z-score is analogous to the dimensionless energy $\varepsilon$, because the $z$-score is a measure of preference or energy difference between the secondary structure states, and therefore, should be proportional to the interaction strength $K$ in a spin lattice system. This is shown in Eq. (1)

$$z \sim \left( \frac{K}{\tau} \right)^{1/2} = \left( \frac{K}{\tau_c (1 + \varepsilon)} \right)^{1/2} \sim \varepsilon$$

(5)

where $\tau$ is an arbitrary temperature, $\tau_c$ the critical temperature, and $\varepsilon$ a variable of small quantity. Ignoring
all constants, the results from de Gennes’ book [8] suggest
that
\[ z \sim \varepsilon \sim z^{-1/\nu} \] (6)
\[ E(z) = E(0) - \text{constant } \varepsilon^{1-\alpha} \] (7)
\[ \alpha = 2 - \nu \text{d} \quad \text{Kadanoff relation} \] (8)
\[ \nu = 3(d + 2) \quad \text{Flory exponent} \] (9)
where \( d \) is the dimension of the space, \( \xi \) the correlation length, and \( E \) the energy of the system. Second, if the density of state \( g(E) \) is assumed to be inversely proportional to energy \( E \) and \( N(z) \) is analogous to \( g(E) \), then using Eqs. (7) and (8), \( N(z) \) is derived as the following:
\[ N(z) \sim g(E) \sim \varepsilon^{-(1-\alpha)} \sim \varepsilon^{-(\nu d - 1)} \sim z^{-(\nu d - 1)} \] (10)
Based on the slopes calculated in the above and Flory’s exponent in Eq. (9), the dimension for the nucleation and collapse stages can be estimated. The nucleation stage gives a dimension of \( d=3.16 \) and the collapse stage \( d=2.09 \). This means that the helix and sheet formation in the nucleation stage is essentially a 3D process and they build-up, as it were, a scaffold or template for the subsequent collapse stage, which would then become a process in 2D space. This is summarized schematically in Fig. 5. The critical nucleus in Fig. 5 provides the surface on to which the rest of the protein residues will collapse.

Further evidence can be found for our argument. Eq. (6) gives the relationship between \( z \) and the correlation length \( \xi \). A higher \( z \) value corresponds to a shorter correlation length. This is consistent with the fact that a higher \( z \) value for a residue corresponds to higher predictive accuracy and hence the more dominant the short-range interaction is in determining the secondary structure of this residue. A more quantitative estimate is possible for helix and sheet. The results of secondary structure prediction show that the size of the correlation window for helix is about 16 residues and that for sheet about 8 residues (data not shown). In polymer physics, the dimensionless correlation length \( \xi \) is defined as the ratio of the persistent length over the total chain length. Here, the correlation length \( \xi \) is approximated by the inverse of the window size for helix and sheet, i.e. \( (a/La)=1/La \), where \( a \) is the unit length of a peptide bond, 0.38 nm, and \( L \) the number of peptide bonds or residues of a typical helix or a sheet strand. In the region of \( z \) values corresponding to the nucleation stage and based on the density of state (or simply the number of residues in each state), the average \( z \) value for a particular \( z \)-value bin can be calculated. Only the ratio between helix and sheet is needed and, for \( z=2.4 \), this ratio is equal to 3.05. Then to satisfy Eq. (6), this ratio should be approximately equal to the ratio of the window sizes for helix and sheet, \( 16/8 \), to the power of \( 1/\nu \). Using Flory’s exponent for a real chain, \( 1/\nu = 1.67 \), and the value of \( (16/8) \) to the power of \( 1/\nu \) is 3.18. This seems to be in reasonable agreement with the experimental value of 3.05.

The density of state derived in Eq. (10) can also be compared with the pair correlation function \( g(r) \) derived for a real chain with \( d=3 \) (Eq. (1.31a) in [8]). By substituting \( z \) for \( r \) using Eq. (6) and Flory’s exponent of 3/5 for \( \nu \), the power law for \( g(z) \) is \( \sim z^{-4/5} \). The exponent \(-4/5\) is in reasonable agreement with the slope (-0.836) found above using Fig. 3. Similarly for the case of \( d=2 \), the power law for \( g(z) \) is \( \sim z^{-1/2} \), again is in reasonable agreement with the slope (-0.533) found above. It is surprising that polymer physics could be applied to polymer chains such as proteins. This should help us study and understand the dynamics of proteins in the future.

4. Conclusions

We start with a nucleation hypothesis. Through designing a new method of secondary structure prediction for proteins, we show that the nucleation residues as hypothesized do exist in native proteins. An important property of nucleation residues is that they exhibit more sequence conservation. Our prediction method is a nearest-neighbor method where only flanking sequences of a residue are considered in making the prediction for the secondary structure of the residue. More importantly, our prediction method calculates a reliability score for each prediction so that higher reliability scores should correspond to a higher predictive accuracy. Using three databases consisting of 1000–4000 proteins, we show that the reliability score correlates well with the predictive accuracy. Based on the nucleation hypothesis, the nucleation residues are expected to correlate with a more conserved sequence, a higher reliability score, a higher predictive accuracy, and finally a more dominant effect contributed by the short-range interactions in determining their secondary structures as opposed to the long-range or tertiary interactions. This is indeed what our results indicate. Therefore, in effect we have created a mechanism to identify nucleation residues.

Further examination shows that there is a clear pattern in the spatial distribution for the 20 amino acids of nucleation residues within proteins, where hydrophobic amino acids are inside and hydrophilic are outside. This is consistent with the widely accepted HP-model for protein folding. It suggests that both the kinetics and thermodynamics are primarily controlled by one and the same set of nucleation residues within a protein sequence.

Then a new physical picture for protein folding begins to emerge: A protein folds initially through the nucleation stage to reduce the effective entropy barrier for folding, and the folding ends with the collapse stage when native-like tertiary structures are formed and the exchange of detailed microscopic conformations occurs and the protein achieves optimal close packing and approaches the global energy minimum. It is shown that both stages exhibit power law
behaviors and could be modeled with critical phenomena as described by de Gennes in his work on polymer statistics [8]. It is shown that the critical exponents estimated based on polymer physics are in reasonable agreement with the experimental values derived from the databases. The fact that the nucleation stage corresponds to a process of a spatial dimension of $d = 3$ and the collapse stage to that of $d = 2$, is consistent with our physical picture for protein folding derived from the nucleation hypothesis as well as most of the experimental observations on protein folding processes.

Acknowledgements

The author is grateful for helpful discussions with both Dr C.C. Lin and Dr K. Huang at MIT during the course of this work. The author thanks Z. Sun and N. Li for help with the secondary structure prediction in this work. The computation for this project was partially performed on the HP-SC45 Sigma-X parallel computer of ITP and ICTS, CAS.

References

[1] D.L. Nelson, M.M. Cox, Lehninger Principles of Biochemistry, fourth ed., W.H. Freeman and Company, New York, 2005.
[2] L. Stryer, Biochemistry, fourth ed., W.H. Freeman and Company, New York, 1995.
[3] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–230.
[4] Kolata, G., Trying to crack the second half of the genetic code. Science 233(1986) 1037–1039.
[5] C. Levinthal, Are there pathways for protein folding?, J. Chim. Phys. 65 (1968) 44–45.
[6] F. Jiang, Prediction of protein secondary structure with a reliability score estimated by local sequence clustering, Protein Eng. 16 (2003) 651–657.
[7] E.E. Lattman, K.M. Fiebig, K.A. Dill, Modeling compact denatured states of proteins, Biochemistry 33 (1994) 6158–6166.
[8] P.G. de Gennes, Scaling Concepts in Polymer Physics, Cornell University Press, Ithaca and London, 1979.
[9] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.
[10] L. Lo Conte, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A. Murzin, SCOP database in 2002: refinements accommodate structural genomics, Nucleic Acids Res. 30 (2002) 264–267.
[11] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[12] J.A. Cuff, G.J. Barton, Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, Proteins 34 (1999) 508–519.
[13] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[14] G.D. Rose, A.R. Gezelowitz, G.J. Lesser, R.H. Lee, M.H. Zehfus, Hydrophobicity of amino acid residues in globular proteins, Science 229 (1985) 834–838.
[15] T. Koshiba, T. Hayashi, I. Miwako, I. Kumagai, T. Ikura, K. Kawano, K. Nitta, K. Kuwajima, Expression of a synthetic gene encoding canine milk lysozyme in Escherichia coli and characterization of the expressed protein, Protein Eng. 12 (1999) 429–435.
[16] R.H. Pain, Mechanisms of Protein Folding, second ed., Oxford University Press, New York, 2000.
[17] Y. Imry, Finite-size rounding of a first-order phase transition, Phys. Rev. 21 (1980) 2042–2043.