Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties.

Alexandre de Brevern, Serge Hazout

Hybrid Protein Model (HPM) : a method to compact protein 3D-structure information and physicochemical properties

Alexandre G. de Brevern, Serge A. Hazout
Equipe de Bioinformatique G´enomique et Mol´eculaire, INSERM U436,
Universit´e Paris 7, case 7113,
2, place Jussieu, 75251 Paris cedex 05, France
debreven@urbb.jussieu.fr, hazout@urbb.jussieu.fr,

Hybrid Protein Model (HPM) : a method to compact protein 3D-structure information and physicochemical properties

Alexandre G. de Brevern, Serge A. Hazout
Equipe de Bioinformatique Génomique et Moléculaire, INSERM U436,
Université Paris 7, case 7113,
2, place Jussieu, 75251 Paris cedex 05, France
debreven@urbb.jussieu.fr, hazout@urbb.jussieu.fr,

Abstract

The transformation of protein 1D-sequence to protein 3D-structure is one of the main difficulties of the structural biology. A structural alphabet had been previously defined from dihedral angles describing the protein backbone as structural information by using an unsupervised classifier. The 16 Protein Blocks (PBs), basis element of the structural alphabet, allows a correct 3D structure approximation [6]. Local prediction had been estimated by a Bayesian approach and shown that sequence information induces strongly the local fold, but stays coarse (prediction rate of 40.7% with one PB, 75.8% with the four most probable PBs).

The Hybrid Protein Model presented in this study learns both sequence and structure of the proteins. The analysis made along the hybrid protein has permitted to appreciate more precisely the spatial location of some types of amino acid residues in the secondary structures and their flanking regions. This study leads to a fuzzy model of dependence between sequence and structure.

key-words : fuzzy model, pattern matching, protein sequence, protein structure, prediction, structural alphabet.

1 Introduction

Proteins fold into a limited number of conformations [9]. The 1-D sequence contains the whole and complete information guiding the protein folding, however we are not able to predict directly the 3D-structure from the sequence [17]. The complexity and the number of physicochemical, kinetic, dynamic and steric parameters does not allow an efficient 3D-structure prediction without the knowledge of 3D-structures of close proteins. A first analysis of protein structures shows the importance of two repetitive secondary structures (α-helix and β-sheet) stabilized by internal bonds. With the variable coils, they constitute a 3-states structural alphabet. The first algorithms of secondary structure prediction were simple and implied statistical methods and information theory like GOR [8]. Their prediction rates were close to 60%. The recent methods including multilayers neural networks and information from homology sequences lead to a maximal accuracy of 75% [20, 21]. However the use of such strategies induces difficulties in the understanding of the factors implied (in particular for the amino acids). The increase of available protein structures permits more precise studies of coil regions [15, 25]. Different teams had developed more complex structural alphabet by taking account of the heterogeneity of backbone protein structures. We could noticed the works of Rooman and co-workers [19] and Fetrow and co-workers [7] based on a limited number (4 to 7) of prototypes of short size (4 to 8 amino acids) strongly dependent of the repetitive secondary structures. Those approaches give poor 3D-structure approximations, however some amino acid distributions seem interesting. Unger and co-workers [24] and Schuchhardt and co-workers [22] have used unsupervised methods (*k-means* method [10] and *Self-Organizing Maps* [13, 14]) to find the most common folds. Hence, the clustering gives 100 distinct groups with a correct 3D-structure approximation, however this important number of cluster is not suited for prediction strategies. Bystroff and Baker [2] had elaborated an interesting method of clusterization and prediction of Structural Blocks of variable lengths (from 3 to 17 residues). The learning step was carried out with both sequences and 3D-structures. The extracted clusters are characteristic of some biological folds, however the structural approximation is only given locally. Recently, we have developed an unsupervised classifier which processes in the same manner than a Self-Organizing Maps [13, 14], and takes account of the main transition in a similar way of the " Hidden Markov Model (HMM) " [18]. The HMM has already been used in that type of study by Camproux and

co-workers [3]. Hence, a 16-states structural alphabet had been defined [6]. The local prediction rate is 34.2% using a Bayesian approach and taking the most probable PB at each site. This rate increases to 40.7% in subdividing the most frequent PBs by an unsupervised method according to their sequence specificity. In fact, the local information is more informative than expected. For instance, 75.8% of the true PBs are found again among the fourth most probable PBs (for the Bayesian approach). From this last remark, we have studied a fuzzy dependence between the protein sequence and the local protein backbone structure. We have developed a learning method called "Hybrid Protein Model" (HPM) to establish this dependence including both sequence and structure in the same observation and keeping the continuity of those observations (i.e. succession of the fragments in the protein). Here, we describe the general principle of this method, then the results obtained on the dependence sequence-structure in terms of amino acid locations. For a better and more precise interpretation, we have studied the correspondence of the hybrid protein with our 16-states structural alphabet. The use of Protein Blocks allows a description closer to the reality than a simple 3-states alphabet which underestimates the importance of variation in coils. In fact, many problems in determination of beginning and end (N- and C-caps) of regular secondary structures (α-helix and β-sheet) exist [4, 5].

2 Material and Methods

2.1 Database

The 3D-protein structure database (Protein Data Bank or PDB [1]) contains at the moment more than 10.000 entries. However many proteins present strong sequence and/or structure similarity. Therefore to tackle the problem induced by the similarity of sequences, only proteins sharing distant sequences are taken in account. Thus, our protein database is composed of 342 proteins which have less than 25% of sequence similarity [12, 11]. Proteins had been cut in fragments of 5 successive residues. The fragments are overlapping, so each protein of length ***L*** is recoding in ***L***-4 fragments, hence the 88 258 residues of the 342 proteins represents 86 980 fragments (i.e. 88 258 - 342 x 4).

Each amino acid had been coded according to three variables: the hydrophobicity [16], the side chain volume [26], and a polarity index (-0.5 assigns to K, R and H +0.5 for D and E, and 0 to the other residues). The two first variables have been normalised in the range [+1.0;-1.0]. The normalised variables are given in Table 1. The 3D-structure associated with the protein backbone is characterised by 5 consecutive residues (the central residue is in position *s* in the protein sequence) is characterised by 8 dihedral angles (ψ_{s-2}, ϕ_{s-1}, ψ_{s-1}, ϕ_s, ψ_s, ϕ_{s+1}, ψ_{s+1}, ϕ_{s+2}) which have been normalised in the range [-1.0;+1.0] after a shift of -360^o for the angles ψ more than 120^o, and of +360^o for the angles ϕ less than -120^o . Thus, each fragment of 5 residues is defined by a vector **V** of 23 component (15 for the sequence and 8 for the structure).

Table 1. Normalised variables: hydrophobicity, volume and polarity index.

amino acid	hydrophobicity	volume	polarity
A	-0.66	+0.40	0.00
R	+0.36	-1.00	-0.50
D	-0.39	-0.78	+0.50
N	-0.36	-0.78	+0.50
C	-0.42	+0.56	0.00
E	-0.06	-0.78	0.00
Q	0.00	-0.78	0.00
G	-1.00	-0.09	0.00
H	+0.11	-0.71	-0.50
I	+0.27	+1.00	0.00
L	+0.27	+0.84	0.00
K	+0.30	-0.87	-0.50
M	+0.23	+0.42	0.00
F	+0.55	+0.62	0.00
P	-0.37	-0.36	0.00
S	-0.65	-0.18	0.00
T	-0.33	-0.16	0.00
W	+1.00	-0.20	0.00
Y	+0.59	-0.29	0.00
V	-0.04	+0.93	0.00

2.2 Hybrid Protein Model (HPM)

In our study, the hybrid protein corresponds to series of *L* fragments of *n* residues, each one is characterised in terms of sequence and structure by a vector of *m* components (here *n*= 5, so *m* = 23). Thus, the hybrid protein is a matrix of dimension *L* x *m*. The principle of the HPM is to learn "at best" the complete database (86980 fragments) by the hybrid protein of *L* vectors. The learning step is similar to a Self-Organizing Map or SOM [13, 14]. However in our case, the training is monodimensionnal and no diffusion of the information along the hybrid protein is performed. In fact, the diffusion is implicitly taken into account since a series of *f* vectors associated with *n*+*f*-1 consecutive residues are presented to the hybrid protein to be learnt (here *f*=5, so 9 consecutive residues are used).

The method relies on three main processes:

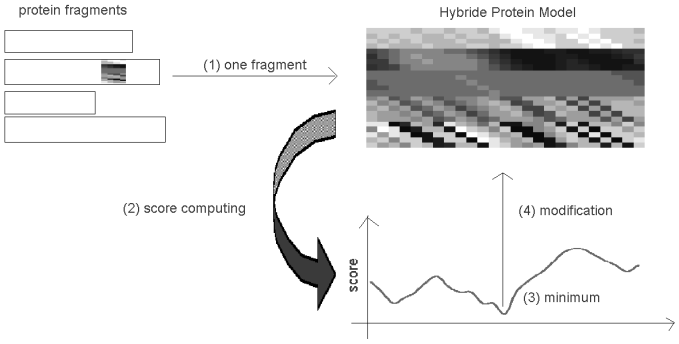

Figure 1. Schema describing the principle of the structure-sequence information learning by the hybrid protein (see the text for more details). (1) Selection of one fragment defined by a sub-matrix. (2) Calculation of a local score between the sub-matrix of this fragment and a region of the hybrid protein. (3) Determination of the optimal position on the hybrid protein in searching for the minimal score. (4) Modification of the local information on the hybrid protein.

(i) Initialisation of the hybrid protein : L vectors are randomly chosen in the structural database (coded in the 23 components).

(ii) Sequential learning of the observation matrices (see Figure 1): (1) one fragment with its environment of length f is taken in the database. It defines a sub-matrix $\boldsymbol{V}$ of f vectors of length m. (2) For every position p of the hybrid protein one score $S(p)$ is computed. The score is an Euclidean distance between the sub-matrices $\boldsymbol{V}$ and $\boldsymbol{W}(p)$ of same length, this latter being located in the hybrid protein. Thus a score profile is established along the hybrid protein. (3) The minimal score S_{min} is found in position $p^* = argmin\{S(p)\}$ associated with the maximal similarity between the protein fragment and a region of the hybrid protein. (4) The sub-matrix $\boldsymbol{W}(p^*)$ is slightly modified to improve its resemblance with the observed one. This transformation is defined by the equation:

$$\boldsymbol{W}(p) \rightarrow \boldsymbol{W}(p) + (\boldsymbol{V} - \boldsymbol{W}(p)).\alpha(n)$$

and

$$\alpha(n) = \alpha_0/(1 + n/N)$$

n is the number of sub-matrices already presented to the hybrid protein, N is the total number of sub-matrices of the structural database and α_0 the initial learning coefficient. The learning coefficient α(n) decreases during the training. The process is reiterated until the complete reading of the database.

(iii) Strengthening of the learning step : It consists in using C times the whole database in the step (ii). A repetitive use of the structural information allows reinforcing the training, so a compacting of close fragments is carried out progressively. The hybrid protein is closed, so the Nth position is contiguous with the first position. With a window of f=5, we have 85 552 sub-matrices $\boldsymbol{V}$ in the structural database (i.e. 86980 - 342 x (f-1)).

2.3 Protein blocks

16 Protein Blocks (labelled from PB*a* to PB*p*) had been defined with the same structural database using an unsupervised classifier which takes account of the transitions existing between protein blocks. These PBs allow a good structural approximation of complete protein 3D-structures [6]. Figure 2 shows fragments superimpositions for the 16 PBs. The protein blocks PB*a* to PB*f* represent the blocks associated with β-sheet, the regular central β is PB*d*, the previous blocks are their N-caps, the following ones their C-caps. As the same, for the blocks associated with α-helix, the block PB*m* corresponds to the regular structures (central part of an right helix) with the blocks PB*k* and PB*l* (N-caps), and PB*n* to PB*p* (C-caps). The last blocks PB*g* to PB*j* are mainly found in coil structures. Table 2 summarizes this information.

Table 2. Protein Blocks (16-states structural alphabet) with the correspondence in the classical 3-states structural alphabet.

Secondary structures	PB labels
N-cap of β-sheet	a, b, c
Central β-sheet	d
C-cap of β-sheet	e, f
Coils	g, h, i, j
N-cap of α-helix	k, l
Central α-helix	m
C-cap of α-helix	n, o, p

3 Results

We present here the main results of the learning step for the protein structures: (i) Description of the hybrid protein in terms of sequence - structure and (ii) Correspondence between the hybrid protein and the protein blocks.

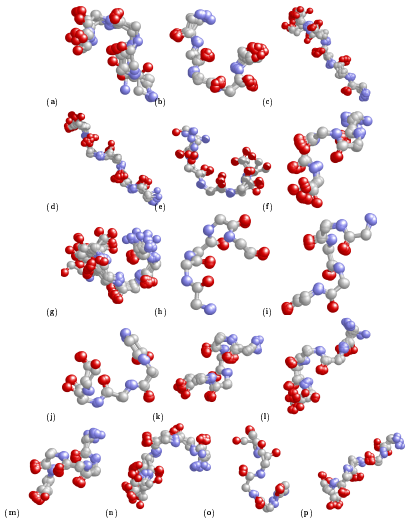

Figure 2. Superimpositions of fragments for the 16 PBs. PB*a* to PB*p* from left to right and from top to bottom .

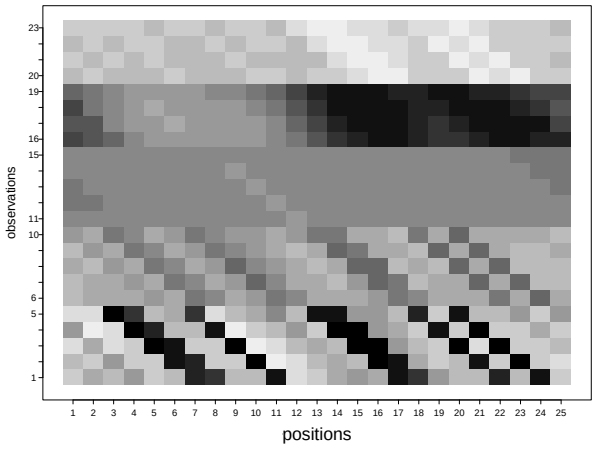

Figure 3. The hybrid protein after the learning step. The hybrid protein is composed of 25 fragments of length 5 (x-axis), characterised by vectors of length 23. Y-axis (for 5 residues): hydrophobicity [1:5], volume [6:10], polarity [11:15] and the dihedral angles ψ [16:19] and ϕ [20:23].

3.1 Description of the hybrid protein in terms of sequence - structure information

Figure 3 shows the hybrid protein obtained with a learning coefficient α_0 = 0,03 for C = 20 cycles and a length L = 25. A trivial observation can be pointed out: the sequentiality of fragments is visible, however we notice that the characteristic vector of fragment in position p does not show exactly the same vector shifted than the fragment in position (p - 1). The variations (in grey scale) of the variables hydrophobicity and dihedral angle ψ show their role in the fragment characterization, and at a lower level, the volume and dihedral angle ϕ. The polarity has a minor part; this can be explained by the limited number of charged residues in regards to the total number; a new estimation of polarity scale is under progress. The 25 patterns present globally different characteristics. A clusterization of those patterns by a hierarchical method shows a significant grouping of two homogeneous distinct groups. The limits between the two groups are the 2^{th} and the 13^{th} positions in the hybrid protein.

After the training, the number of fragments located in each position has been computed. The observations are evenly distributed along the hybrid protein, the numbers varying between 2950 and 4500 observations.

3.2 Correspondence between the hybrid protein and the protein blocks

Figure 4 gives the amino acid composition of the central residue for the 25 fragments on the left side, and the relative frequencies of the fragments for each protein block. The 16 protein blocks [6] allow a more precise description of the 3D structures than the secondary structure. For a better visualisation, only the groups with a frequency more than 4% (i.e. 1/L) are shown in the figure. Only 15.6% of the fragments are not taken into account with this rule.

The analysis of the results allows one to point out:

(i) A strong dependence between the hybrid protein fragments and the protein blocks; from position 2 to 13, the blocks associated with the α-helix and their flanking regions, and from 14 to 25 (and position 1), the β-sheets, their flanking zones and the coils. The coils are found at positions 1, 12, 13, [17:19] and mainly in [22:25]. We notice an overexpression of glycine (G) and proline (P) in those zones in association with blocks PB*h*, PB*i* and PB*j*.

(ii) The sequentiality of the protein blocks and protein hybrid fragments (see Figure 4b). We observe two tendencies in the β-sheets, one for the positions 13 to 19, the other ones from 20 to 25. The first β-sheet has two following hydrophobic positions (positions 15 and 16) and a high propensity of PB*d* (central β-sheet). It corresponds to an internal β-sheet. The second β-sheet has two hydrophobic

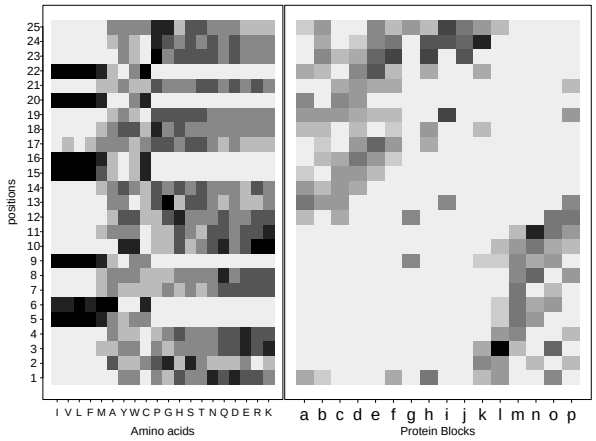

Figure 4. Correspondence between the hybrid protein and the protein blocks. (a) Left side : Amino acid locations along the hybrid protein (for the central site of the protein fragments). (b) Right side : Fragment occurrences according to the position i along the hybrid protein ($1 \leq i \leq 25$) and the type of PB (from a to p).

positions (20 and 22) separated by an hydrophilic position (position 21) defines a specific β-sheet. A part of the β-sheet is exposed (accessible to the solvent) at the protein surface and the other one is buried.

(iii) An important role of the physicochemical properties in the 3D-structure: In fact, some PBs are found in a limited number of positions. For instance, PB*l* (N-cap of α-helix) is located in positions 1 to 4, positions with an over-expression of charged residues, PB*m* (central α-helix) in positions 4 to 11, where hydrophobic residues (I, V, L, F, M et partially A). We could notice that the hydrophobic motif (i, i+1, i+4) is found again and characterises the buried sites. The central positions (5, 6, 15, 16, 20 et 22) express a hydrophobic state. The cysteine (i.e. which creates the most stable contact) is observed in the same places but with a lower intensity. The implication of charged residues is less precise and is not as important as expected in the flanking regions (N- and C-caps) of α-helices and β-sheets.

4 Discussion and conclusion

The PBs location as seen in Fig. 4 shows strong tendencies of regrouping which could be compared to the classical 3-states structural alphabet (α-helix, β-sheet and coils). A simple counting of the 3 structures gives an information, as expected correlated with the one obtained by the hybrid protein. However, this information does not take into account the sequentiality. A 16-states alphabet allows a more precise structural information. For instance, the N- and C-caps are located at distinct positions and so have distinct amino acid preferences, the hybrid protein gives an assessing of those propensities. The N- and C-caps could begin at different positions. The beginning of the α-helix could be noticed (mainly around the position 3), the same for the C-cap of α-helix which goes from positions 8 to 11. With the β-sheets, the learning had permitted the obtaining of two types of β-sheet (positions [14:17] and [19:23]) of different lengths and distinct amino acid compositions. The hybrid protein allows to take account of the length heterogeneity of the repetitive structures. This fact is not easily perceptible with the three-states alphabet.

We have tested different values of parameters: (i)The choice of a length L=25 allows a good assignation for most of the fragments in the hybrid protein (i.e. 84.4% of the fragments composing the structural database, for a threshold of 4% ($1/L$) per site). A longer hybrid protein had decreased this rate. A shorter hybrid protein did not allow the distinction of the two types of β-sheets. (ii) By taking a high α_0 value, the training is not sensitive to the definition of the initial hybrid protein. We have observed similar clusterization for different initial α_0 values. (iii) We have chosen a sufficient fragment length (f=5, e.g. 9 Cα) to insure the structure continuity along the protein. They are long fragments from a structural point of view (longer than short regular structures).

The hybrid protein allows the compaction of both sequence and structures in a finite number of states where the two information types are combined. A structural alphabet of the 16 PBs constitutes an efficient tool for analysing the protein structure. In conclusion, the correspondence between the fragment series in the hybrid protein and different types of protein blocks permits to propose a fuzzy concept of relationship sequence-structure. It seems true that an amino acid strings is associated with a set of structural patterns (i.e. PBs series) and conversely. It implies that in a prediction method "sequence to structure", some protein blocks must be considered as equivalent. We could noticed that Simmons and co-workers [23] had already used Structural Blocks to improve *ab initio* modelling. So this type of concept might be useful in the molecular modelling as the *threading* technics or in structural homology method. A work under progress concerns the use of this sequence-structure table as a tool of assessment of a protein backbone modelling.

References

[1] F. Bernstein, T. Koetzle, G. Williams, E. Meyer, M. Brice,

J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, 112:535–540, 1977.

[2] C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motif. *J Mol Biol*, 281:565–577, 1998.

[3] A. Camproux, P. Tuffery, J. Chevrolat, J. Boisvieux, and S. Hazout. Hidden markov model approach for identifying the modular framework of the protein backbone. *Protein Eng*, 12(12):1063–1073, 1999.

[4] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, and J. Mornon. Comparaison of three algorithms for the assignement of secondary structure in proteins: the advantages of a concenssus assignement. *Protein Eng.*, 6:377–382, 1993.

[5] J. Cuff and G. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34:508–519, 1999.

[6] A. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks. *Proteins*, forthcoming 2000.

[7] J. Fetrow, M. Palumbo, and R. Berg. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins*, 27:249–271, 1997.

[8] J. Garnier, D. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular protein. *J Mol Biol*, 120:97–120, 1978.

[9] S. Govindarajan, R. Recabarren, and R. Goldstein. Estimating the total number of protein folds. *Proteins*, 35:408–414, 1999.

[10] J. Hartigan and M. Wong. k-means. *Applied Statistics*, 28(1):100–115, 1979.

[11] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Prot Sci*, 3:522–524, 1994.

[12] U. Hobohm, F. Scharf, S. R, and C. Sander. Selection of a representative set of structures from the brookhaven protein databank. *Prot Sci*, 1:409–417, 1992.

[13] T. Kohonen. Learning vector quantization. *Neural Networks*, 1(suppl. 1):303, 1989.

[14] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, Germany, 1997.

[15] J. Kwasigroch, J. Chomilier, and J. Mornon. A global taxonomy of loops in globular proteins. *J Mol Biol*, 259:855–872, 1996.

[16] J. Kyte and R. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, May 1982.

[17] C. Levinthal. Molecular model-building by computer. *Sci Am*, 214:42–52, 1966.

[18] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77:257–285, 1989.

[19] M. Rooman, J. Rodriguez, and S. Wodak. Automatic definition of reccurent local structure motifs in proteins. *J Mol Biol*, 213:327–336, 1990.

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

[21] A. Salamov and V. Solovyev. Protein secondary structure prediction using local alignenments. *J Mol Biol*, 268:31–36, 1997.

[22] J. Schuchhardt, G. Schneider, J. Reichelt, D. Schomburg, and P. Wrede. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng*, 9(10):833–842, 1996.

[23] K. Simons, R. Bonneau, I. Ruczinski, and D. Baker. Ab initio protein structure prediction of casp 3 targets using rosetta. *Proteins*, 34(suppl. 3):171–176, 1999.

[24] R. Unger, D. Harel, W. S, and S. JL. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5:355–373, 1989.

[25] J. Wodjick, J. Mornon, and J. Chomilier. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol*, 289:1469–1490, 1999.

[26] A. Zamyatin. Protein volume in solution. *Prog Biophys Mol Biol*, 24:107–123, 1972.