Theor Chem Acc (1999) 101:16–20
DOI 10.1007/s002149800m67

Theoretical
Chemistry Accounts

Regular article

Prediction of protein structure using a knowledge-based off-lattice united-residue force field and global optimization methods*

Adam Liwo[1,2], Jarosław Pillardy[2], Rajmund Kaźmierkiewicz[1], Ryszard J. Wawak[2], Małgorzata Groth[1], Cezary Czaplewski[1], Stanisaw Ołdziej[1], Harold A. Scheraga[2]

[1] Faculty of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland
[2] Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA

Received: 24 April 1998 / Accepted: 4 August 1998 / Published online: 2 November 1998

Abstract. A united-residue model of polypeptide chains developed in our laboratories with united side-chains and united peptide groups as interaction sites is presented. The model is designed to work in continuous space; hence efficient global-optimization methods can be applied. In this work, we adopted the distance-scaling method that is based on continuous deformation of the original rugged energy hypersurface to obtain a smoothed surface. The method has been applied successfully to predict the structures of simple motifs, such as the three-helix bundle structure of the 10-58 fragment of staphylococcal protein A in de novo folding simulations and more complicated motifs in inverse-folding simulations.

Key words: Protein structure prediction – Mean-field potential – United-residue representation of polypeptide chains – Global optimization

1 Introduction

Prediction of protein structure from amino acid sequences is still an unsolved problem of contemporary molecular biology [1]. Its significance lies primarily in the imbalance between the huge number (tens of thousand) of new protein sequences that are discovered yearly and only about 200 3D structures that have been solved during this period. Moreover, knowing the unique rules, according to which a sequence is transformed into a 3D structure, will enable the prediction of the effect of mutations on the stability of the structure of vital proteins, which has great significance for health sciences (e.g. for cancer research). Apart from these practical reasons, research on protein-structure prediction continues to contribute a great deal to studies of the mechanisms of protein folding.

There are three classes of approaches to the structure-prediction problem: sequence-homology methods, methods based on energetic criteria, and threading methods. In the first method, the unknown structure is constructed based on known structural motifs whose amino acid sequences are similar to the sequence studied, taking advantage of the empirical relationship between sequence and 3D structure [2–5]. The methods of the second group [1, 6, 7] are based on the thermodynamic hypothesis formulated by Anfinsen [8], according to which the native structure of a protein is the global minimum of its free energy under given conditions. Structure prediction is therefore achieved by a global-minimum search of an appropriate free-energy function; this is called the ab initio or de novo approach. The threading methods can be placed between these two approaches: they use the energy (or energy-like) functions to distinguish the native structure from alternative structures, but the unknown sequence is superposed on the structural motifs chosen from a database of known protein structures [9].

In this report, we describe a de novo method for protein-structure prediction that is under development in our laboratories. In order to reduce the complexity of the problem, a united-residue representation of polypeptide chains is used. Because the model works in continuous space, efficient global-optimization methods can be applied in a global-minimum search.

2 Methods

2.1. Representation of polypeptide chains and interaction scheme

In our model [10–12] a polypeptide chain is represented by a sequence of α-carbon (C^{α}) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p)

* Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: H.A. Scheraga

located in the middle between the consecutive C^α. Only the united p and united SC serve as interaction sites, the C^α assisting in the definition of the geometry (Fig. 1). All the virtual bond lengths (i.e., C^α—C^α and C^α—SC) are fixed; the C^α—C^α distance is taken as 3.8 Å which corresponds to trans p, while the side-chain (α_{SC} and β_{SC}), as well the virtual-bond angles θ can vary.

The energy of the virtual-bond chain is expressed by Eq. (1).

$$U = \sum_{i<j} U_{SC_iSC_j} + \sum_{i\neq j} U_{SC_ip_j} + w_{el} \sum_{i<j-1} U_{p_ip_j} + w_{tor} \sum_i U_{tor}(\gamma_i) + w_{loc} \sum_i [U_b(\theta_i) + U_{rot}(\alpha_{SC_i}, \beta_{SC_i})] + w_{corr} U_{corr} \quad (1)$$

The term $U_{SC_iSC_j}$ pertains to the mean free energy of the hydrophobic (hydrophilic) interactions between the SC. It, therefore, contains implicitly the contributions coming from the interactions with the solvent. The terms with $U_{SC_ip_j}$ denote the excluded-volume potential of the SC – p interactions. The p interaction potential ($U_{p_ip_j}$) accounts mainly for the electrostatic interactions between them or, in other words, for their tendency to form backbone hydrogen bonds. U_{tor}, U_b, and U_{rot} denote the energies of virtual-dihedral angle torsions, virtual-angle bending, and SC rotamers; these terms reflect the local propensities of the polypeptide chain. Finally, the multibody (or cooperative) term U_{corr} arises from the fact that details of the all-atom chain are lost when converting it into the simplified chain. Mathematically, it can be expressed by averaging the energy over some "less important" degrees of freedom; this gives rise not only to a pairwise potential, but also to higher-order terms [13]. For the functional form of these energy terms, the reader is referred to the original papers [10–14]. The ws denote relative weights of the respective energy terms; they are discussed later.

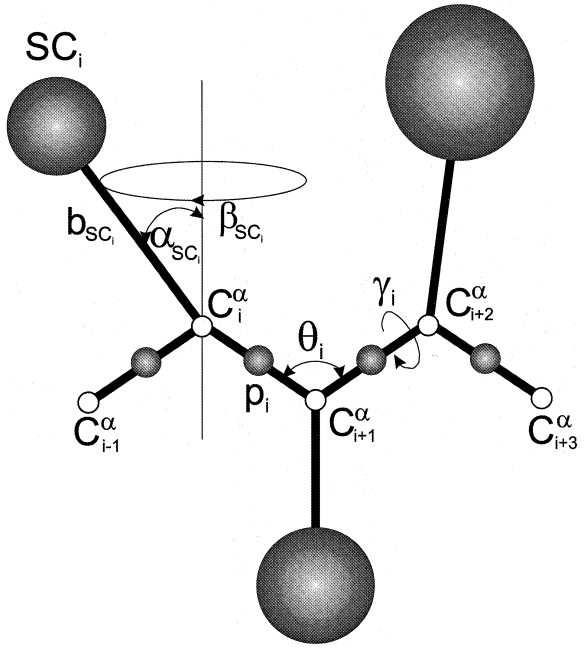

Fig. 1. United-residue representation of a polypeptide chain. The interaction sites are side-chain (*SC*) centroids of different sizes and peptide-bond centers (*p*) indicated by *shaded circles*, while the α-carbon atoms (C^α) (*small empty circles*) are introduced only to assist in defining the geometry. The virtual C^α—C^α bonds have a fixed length of 3.8 Å, corresponding to a *trans* peptide group; the virtual-bond (θ) and dihedral (γ) angles are variable. Each SC is attached to the corresponding C^α with a fixed "bond length", b_{SC_i}, variable "bond angle", α_{SC_i}, formed by SC_i and the bisector of the angle defined by C^α_{i-1}, C^α_i, and C^α_{i+1}, and with a variable "dihedral angle" β_{SC_i} of counterclockwise rotation about the bisector, starting from the right side of the C^α_{i-1}, C^α_i, C^α_{i+1} frame

2.2. Parameterization of the force field

As in other work on structure-derived protein potentials [15–18], the SC (U_{SCSC}) and the components of the local-interaction potential (U_b and U_{rot}) of our united-residue force field have been parameterized based on distribution and correlation functions determined from a set of 195 high-resolution non-homologous structures from the Protein Data Bank (PDB) [19]. This approach assumes that average interactions can be described with sufficient accuracy by using the potential of mean force, $W(\mathbf{X})$, $\mathbf{X}$ denoting the degrees of freedom of the subsystem considered. The degrees of freedom are related directly to the corresponding distribution functions, $\rho(\mathbf{X})$:

$$\rho(\mathbf{X}) = \rho_\circ(\mathbf{X}) \exp[-\beta W(\mathbf{X})] \ , \quad (2)$$

where $\rho_\circ(\mathbf{X})$ is a known reference distribution function (e.g., the distribution function of non-interacting side chains tethered to the backbone).

The components of an empirical force field can therefore be parameterized by fitting the theoretically calculated (Eq. 2) to the corresponding experimental distribution functions; this approach has been implemented in our work [12, 14]. The peptide-group interaction potential $U_{p_ip_j}$ and the virtual-torsional potential U_{tor} were parameterized [10, 11] by averaging the all-atom ECEPP/2 [20, 21] potential.

The final stage of parameterization was to calculate the relative weights of the components of the force field so as to maximize the energy gap between the native structure and the lowest-in-energy non-native structure. We have developed a method [14] based on the approach of Wolynes and coworkers [22], Shakhnovich and coworkers [23], and Hao and Scheraga [24, 25], which is directed at achieving as negative a value as possible for the so-called Z-score function (Eq. 3).

$$Z = \frac{E_\circ - 1/N \sum_{i=1}^N E_i}{\sqrt{1/N \sum_{i=1}^N E_i^2 - 1/N^2 (\sum_{i=1}^N E_i)^2}} \ , \quad (3)$$

where N is the number of non-native conformations, $E_\circ$ is the energy of the native conformation, and E_i is the energy of the ith non-native conformation.

The more negative the Z-score values, the more the native structure is distinguished from non-native ones. To optimize the weights, we chose the phosphocarrier protein from *Streptococcus faecalis* (1PTF; 87 amino-acid residues; 1.6 Å resolution). Its structure is shown in Fig. 2. The set of non-native conformations was constructed by superposing the 1PTF sequence on randomly chosen polypeptide-chain fragments of appropriate length from the PDB and minimizing the energies of the chains; this can be considered a version of the inverse folding approach. The procedure is described in detail in the original paper [14].

2.3. Search of the conformational space

In order to search the conformational space of the simplified chain, we used two protocols. The first one is a direct implementation of the Monte Carlo with minimization (MCM) method of Li and Scheraga [26, 27]. In brief, it can be summarized as follows:

1. Choose an arbitrary starting conformation.
2. Minimize the energy; let the geometric parameters of the resulting conformation be contained in the vector $\Gamma_\circ$ and the corresponding energy be $U_\circ$.
3. Perturb $\Gamma_\circ$ according to a predetermined scheme.
4. Carry out energy minimization, obtaining the conformation Γ_1 and energy U_1.
5. If neither U_1 nor Γ_1 differs by more than preassigned cut-off values, discard it and repeat the process beginning at step 3; otherwise apply a Metropolis test [28] in order to accept or reject the conformation.
6. If the new conformation is accepted, substitute Γ_1 for $\Gamma_\circ$, and U_1 for $U_\circ$, and repeat from step 3.

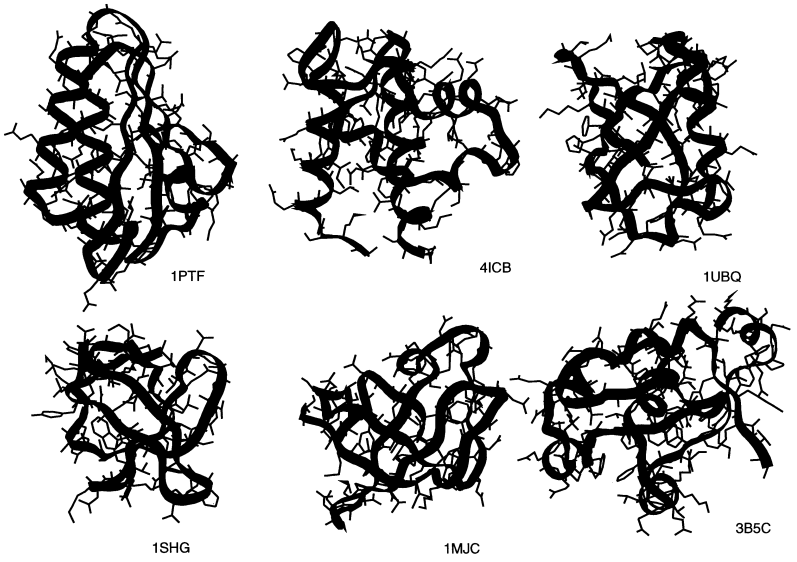

Fig. 2. The structures of the proteins used in inverse-folding calculations: histidine-containing phosphocarrier protein (*1PTF*), calcium-binding protein (*4ICB*), ubiquitin (*1UBQ*), α-spectrin (*1SHG*), major cold-shock protein (*1MJC*), and cytochrome b5c (*3B5C*)

Table 1. Summary of threading calculations with weights determined using the phosphocarrier protein (1PTF) (data from Ref. [14])

Protein[a]	N[b]	Type	Cofactor	ΔE_{nat} (kcal/mol)	*Z*-score	RMSD[c] (Å)
4ICB	76	α	Ca^{2+}	−24.6	−4.84	4.5
1UBQ	76	$\beta+\alpha$	None	−15.5	−3.29	3.1
3B5C	85	$\alpha+\beta$	heme,Fe^{2+}	−13.5	−3.56	3.1
1SHG	57	β	None	−5.2	−3.06	2.5
1MJC	69	β	None	−3.9	−3.40	2.5

[a] See Fig. 2 for the names of these proteins
[b] The number of amino acid residues
[c] Residual-mean-square (RMS) deviation from the native structure

7. Iterate steps 3–7, until the requested number of accepted conformations is obtained.

The second protocol, which is still under development [29], uses the MCM method in combination with the distance-scaling method (DSM) [30] which smoothes the energy surface by applying the following transformation to the site-site distances (other smoothing methods, such as the diffusion equation method [31, 32] can also be implemented here):

$$\tilde{r}_{ij} = \frac{r_{ij} + a r_{ij}^{o}}{1 + a} \quad , \tag{4}$$

where r_{ij}^{o} usually is the position of the minimum in the pairwise potential or an arbitrary large distance, if the potential does not have a minimum. The greater the value of the parameter a, the flatter the transformed energy surface, which facilitates the global-minimum search. The original energy surface is obtained with $a = 0$. The algorithm that combines the MCM search with the DSM smoothing technique is outlined below; it has been implemented in our previous work in the prediction of crystal structures of small molecules [33, 34].

1. Choose a starting value of a; this value should be large enough to smooth the energy surface considerably. Usually a value between 8 and 15 is chosen.

2. Carry out an MCM search of the deformed energy surface. Store N (N being a pre-determined number) of the resulting low-energy conformations for the next steps; usually N varies from 2 to 30. Larger values of N result in better efficiency of the search, but lead to greater CPU usage.

3. Decrease the deformation parameter a according to the assumed schedule. Best results were obtained using the exponential formula: $\log a_{n+1} = \log a_n - \Delta$, where n is the deformation-reversing step and Δ is a pre-assigned increment, usually 0.1 or 0.2. Carry out MCM searches, starting from the low-energy conformations obtained in the preceding deformation step. Select N lowest-energy conformations for the next steps.

4. Iterate step 3 until a reaches 0 (this value corresponding to the original energy surface). The result of the procedure is the lowest-energy conformation obtained with $a = 0$.

Inclusion of the MCM search in step 3 of the algorithm greatly increases its efficiency, although the procedure can also work without this. The efficiency is particularly increased when a more extensive MCM search is carried out in the final stage (on the undeformed energy surface). Searches on the highly deformed surface have less impact on the efficiency of the procedure.

So far, the method has been applied to model polyalanine chains [29]; at the present stage, the method can handle up to 60-residue chains.

3. Results

3.1. Inverse-folding results

Using the weights determined from the inverse-folding calculations on the phosphocarrier protein (1PTF), we checked the ability of the potential to locate the native structures of other proteins correctly, using the inverse-folding approach [14]. In these calculations, the force field did not include the correlation term U_{corr}. Table 1 summarizes the results of these calculations for a number of monomeric proteins of length exceeding 50 amino acid residues, and the structures are shown in Fig. 2. As shown, in each case the native structure is the

lowest in energy and is separated from non-native structures by a significant energy gap. This means that weights obtained by calibrating the energy function using the phosphocarrier protein (1PTF) are also relevant for other proteins. It should be noted that none of the above proteins was used in parameterization of the potential.

Use of energy minimization in our threading calculations provides the possibility that the procedure will find a structure close to the native pattern of the target protein, even if the structural fragments from the data base are distant from its native structure. The results of threading-with-minimization calculations for the 10-58 fragment of the B-domain of staphylococcal protein A are summarized in Table 2. The native pattern of protein A was not present in the data base. As shown, all but one of the five lowest-energy patterns found in the data base are close to the native structure of protein A. It is also important to note that the residual mean-square (RMS) deviations of energy-minimized structures 4 and 5 from the corresponding starting structure are as large as 7.4 and 9.5 Å, respectively, while, at the same time, these structures approached the native structure of protein A quite closely (the RMS deviations being 4.1 and 4.4 Å, respectively).

Table 2. Summary of the results of the calculations on the 10–58 fragment of staphylococcal protein A (data from Ref. [35])

Structure[a]	Start[b]	Energy (kcal/mol)	RMSDp[c] (Å)	RMSDn[d] (Å)	%NC[e]
1BAB:D	23	−146.5	4.4	3.8	55
1CSC	144	−143.7	5.1	5.8	43
1ECA	2	−143.6	9.1	9.5	43
2HMZ:C	48	−143.3	7.4	4.1	48
1CPC:B	1	−141.7	9.5	4.4	52
Native	10	−149.7	2.5	2.5	89

[a] Four-digit code of the Protein Data Base (PDB) entry followed by the chain code, if applicable
[b] The residue of the PDB structure onto which the first residue of protein A was superposed
[c] RMS deviation of the energy-minimized structure from the original PDB pattern
[d] RMS deviation from the NMR structure of protein A
[e] Percentage of native contacts (NC)

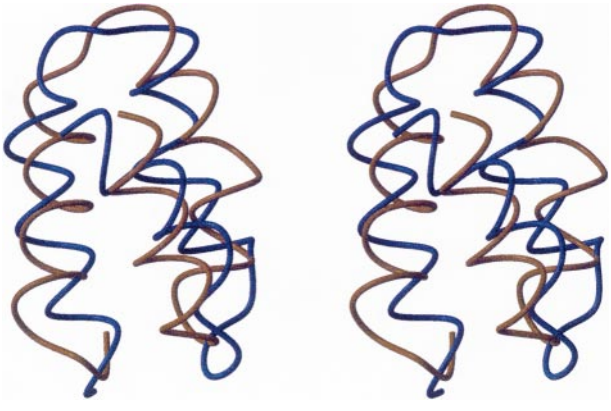

Fig. 3. The ribbon representation of the lowest-energy conformation of the 10–58 fragment of the B-domain of the staphylococcal protein A (*orange*) superposed on the native structure crystal structure (*blue*). The residual-mean-square deviation from the C^{α} trace is 3.5 Å

3.2. De novo folding of the 10–58 fragment of the B-domain of staphylococcal protein A

Using the MCM procedure described in Sect. 3 we attempted to carry out a de novo prediction of the native structure of the 10–58 fragment of the B-domain of staphylococcal protein A. The force field included the correlation term (Eq. 1); failure to do so leads to too low a stability of the secondary structure [13]. Inclusion of cooperativity terms is also advisable in threading-with-minimization calculations, because it better differentiates, in terms of energy, the native structure from the non-native structures, owing to secondary-structure promotion. In order to compare the conformations in step 5 of the MCM procedure (cf. Sect. 2.3), we used the average difference in virtual-bond dihedral angles γ defined by Eq. (10) of Ref. [11], the cut-off value being 10°. Four runs starting from randomly generated structures subject to the condition of non-overlap were carried out; each run was terminated after 1,000 accepted conformations. Two major structures were obtained: the native-like structure of protein A [with a RMS deviation from the native structure of 3.5 Å (calculated based on C^{α} atoms) and 60% of native contacts (Fig. 3)] and its mirror image (with an RMS deviation from the native structure of 9 Å). The native-like structure was marginally stable (by 1 kcal/mol) with respect to the alternative structure. This result is probably a consequence of the fact that the weights of the energy terms were determined by using inverse-folding calculations (cf. Sect. 2). The structures from the PDB used in such calibration of the force field usually correspond to favorable local interactions and also consist of regular patterns. It can therefore be supposed that, while they can serve to determine the relative weights of the hydrophobic and electrostatic terms, the weights of the local and correlation terms will be estimated poorly. We are now working on extending the determination of the weights of the energy terms (Eq. 1) to structures outside the PDB.

Acknowledgements. This work was supported by grant 3 T09A 036 10 from the Polish State Committee for Scientific Research (KBN) to A.L., by grant GM-14312 from the National Institute of General Medical Sciences to H.A.S., and by grant MCB95-13167 from the National Science Foundation to H.A.S. The computations were carried out at the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, Poland, the Cornell Theory Center in Ithaca, N.Y., and the Interdisciplinary Center for Mathematical Modeling (ICM) in Warsaw, Poland.

References

1. Scheraga HA (1996) Biophys Chem 59:329
2. Jones TA, Thirup S (1986) EMBO J 5:819
3. Clark DA, Shirazi J, Rawlings CJ (1991) Protein Eng 4:751
4. Rooman MJ, Wodak SJ (1992) Biochemistry 31:10239
5. Johnson MS, Overington JP, Blundell TL (1993) J Mol Biol 231:735

6. Scheraga HA (1992) Int J Quantum Chem 42:1529
7. Vsquez M. Némethy G. Scheraga HA (1994) Chem Rev 94:2183
8. Anfinsen CB (1973) Science 181:223
9. Fischer D, Rice D, Bowie JU, Eisenberg D (1996) FASEB J 10:126
10. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA (1993) Protein Sci 2:1697
11. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA (1993) Protein Sci 2:1715
12. Liwo A, Ołdziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA (1997) J Comput Chem 18:849
13. Liwo A, Kaźmierkiewicz R, Czaplewski C, Groth M, Ołdziej S, Wawak RJ, Rackovsky S, Pincus MR, Scheraga HA (1998) J Comput Chem 19:259
14. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Ołdziej S, Scheraga HA (1997) J Comput Chem 18:874
15. Sippl MJ (1993) J Comput Aided Mol Des 7:473
16. Godzik A, Koliński A, Skolnick J (1993) J Comput Aided Mol Des 7:397
17. Hinds DA, Levitt M (1994) J Mol Biol 243:668
18. Crippen GM (1996) J Mol Biol 260:467
19. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) J Mol Biol 112:535
20. Momany FA, McGuire RF, Burgess AW, Scheraga HA (1975) J Phys Chem 79:2361
21. Némethy G, Pottle MS, Scheraga HA (1983) J Phys Chem 87:1883
22. Goldstein RA, Luthey-Schulten ZA, Wolynes PG (1992) Proc Natl Acad Sci USA 89:9029
23. Šali A, Shakhnovich E, Karplus M (1994) Nature 369:248
24. Hao M-H, Scheraga HA (1994) J Phys Chem 98:9882
25. Hao M-H, Scheraga HA (1996) Proc Natl Acad Sci USA 93:4984
26. Li Z, Scheraga HA (1987) Proc Natl Acad Sci USA 84:6611
27. Li Z, Scheraga HA (1988) J Mol Struct (Theochem) 179:333
28. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) J Chem Phys 21:1087
29. Pillardy J, Liwo A, Groth M, Rackovsky S, Scheraga HA (1998) (unpublished)
30. Pillardy J, Olszewski KA, Piela L (1992) J Mol Struct 270:277
31. Piela L, Kostrowicki J, Scheraga HA (1989) J Phys Chem 93:3339
32. Kostrowicki J, Scheraga HA (1992) J Phys Chem 96:7442
33. Wawak RJ, Gibson KD, Liwo A, Scheraga HA (1996) Proc Natl Acad Sci USA 93:1743
34. Wawak RJ, Pillardy J, Liwo A, Gibson KD, Scheraga HA (1998) J Phys Chem A 102:2904
35. Liwo A, Ołdziej S, Czaplewski C, Groth M, Kaźmierkiewicz R, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA (1996) International Symposium on Theoretical and Experimental Aspects of Protein Folding, San Luis, Argentina, 17–21 June, 1996