Chaos, Solitons & Fractals Vol. 1, No. 4, pp. 375 - 382, 1991
Printed in Great Britain

0960-0779/91$3.00+.00
Pergamon Press plc

Protein Folding and Deterministic Chaos: Limits of Protein Folding Simulations and Calculations

Gerald Böhm

Institut für Biophysik und Physikalische Biochemie
Universität Regensburg, D-8400 Regensburg, F.R.G.

Abstract - The aim of the present work is to suggest that protein folding is a highly complex process which generally cannot be simulated on digital computers. This limitation is not due to the availability of *computing ressources* or exact *force field parameters*, as it has been suggested previously; it is obviously impossible to quantify parameter(s) for any deterministic algorithm with sufficient accuracy to describe the dynamics of protein folding. Molecular dynamics simulations on crambin, a small protein with 46 amino acids whose three-dimensional structure is known, suggest the native state to be a *'fixed attractor'*. The results show that any *ab initio* calculation of protein structure must fail if the folding process of a protein is controlled by a *kinetic* process, i.e. when the native state is the *kinetically accessible minimum* on the energy hyperspace but not the *thermodynamically* possible global minimum. In this case only non-dynamical methods like *pattern recognition* or *database processing* (knowledge-based approaches) can provide a reasonable three-dimensional structure. Novel computational methods like *genetic algorithms* and *neural network* methods may be more valuable for the design and description of protein structures than the traditional force-field based algorithmic methods used to date. Knowledge-based modelling is therefore the most promising method to date to deduce the structure of an unknown protein from the sequence information solely.

INTRODUCTION

The protein folding problem is sometimes considered to be the *'second part of the genetic code'*[1], defining the transition from the translated protein sequence to the native tertiary structure. The information required for this process is intrinsically embedded in the protein sequence and the solvent conditions, preferrably the conditions *in vivo*. A 'folding matrix', e.g. information from external sources, is not neccessary; however, it has been suggested that *chaperones* may be of relevance for *in vivo* folding[2].

All attempts to extract the information for the sequence-structure relationship have failed so far. The most promising method to deduce the structure of an unknown protein is *knowledge-based* (*'comparative'*) modelling[3]; here, the generation of structures is guided by known structures from homologous proteins.

Other methods which have been published so far do either lack general applicability[4], or the calculation was directed by some knowledge of the native structure[5]. Also, for a correct description of functional processes of the protein, the precision for atom positions required to describe at least the active site and/or ligand-binding parts must be better than 1 Å.

In this work it is sugested that protein folding is a process of highly *deterministic chaos*, as it has been shown for a multitude of natural events. The term 'deterministic chaos' was first suggested during investigations on climate events[6]; until today, many non-linear dynamical processes in nature have been shown to be unpredictable, although the fundamental principles underlying those events are known[7].

This is mainly due to the problem of describing initial conditions as well as parameters for an algorithm in terms of our mathematics; in the axiomatic system of scientific explanation which is currently used, most descriptions of events are based on decimal numbers with limited precision. On the pathways of dynamical events, the errors intrinsically included in any decimal number will add up, until finally the error (that was negligible in the beginning) will be of greater magnitude than the corresponding value.

One way to circumvent the algorithmic problem is to use a *neural network* (NN)[8]. This method is based on the fact that patterns can be learned by special programs (usually in a training- and a recall-phase). The basis for the learning are layers of '*neurons*' with different weights ('*synaptic strength*'). New patterns are associated with the stored information.

Another new method in computer algebra is the *genetic algorithm* (GA). This development has its parallels in biological evolution. GAs find the solution to a problem by analyzing the feedback to repeated attempts of solutions. The attempts toward the solution are called *genes* - a sequence of information located in the problem space. GA have no *a priori* knowledge of the problem space or the environmental conditions; this is scanned during the evolution of the solution. Similarly, '*fuzzy logic*' methods may circumvent the problem of fixed rules for dynamics and energy minimization[9].

MATERIALS AND METHODS

The protein used for the calculations described later is crambin, a small plant seed protein with 46 amino acids from *Crambe abyssinica*. It has two short helical regions, two β-strands, and three disulfide bridges; it is one of the smallest stable proteins with defined secondary structural elements whose three-dimensional structure is known so far. The resolution of the crystal structure is 1.0 Å.

All molecular dynamics simulation described in this work were performed with the program DISCOVER (Biosym Inc., San Diego, U.S.A.), release 2.6 for the IRIS 4D series (Silicon Graphics Inc., Mountain View, U.S.A.), and DISCOVER release 2.5 for the CRAY Y-MP 4/432. Both machines were running under UNIX derivates (IRIX 3.3.2 and UNICOS 5.2, respectively). The numerical differences for identical calculations performed on both machines were determined to be below 0.5 %; these differences are based mainly on the internal data representation and the processing methods used. A simulation of Crambin on the IRIS 4D/70, equipped with a 12.5 MHz MIPS R3000 RISC processor, took about 12,000 seconds CPU time per picosecond of simulation, whereas the CRAY Y-MP 4/432, equipped with 4 processors, took about 60-80 seconds per simulated picosecond.

Simulations were performed mainly under *in vacuo* conditions. All hydrogens were present in the structure. The cutoff distance for intramolecular interactions between atoms was 18.0 Å, with 2.0 Å switching

distance. A *leapfrog algorithm* for the startup of the dynamics was used. Initial equilibrium of the dynamics was performed over 5.0 ps; the simulations took 80 ps or 100 ps, respectively, with an underlying time step of 1 fs. Structures were collected each 0.1 ps.

For comparison reasons, one calculation was performed under identical coditions with respect to the above calculations, but with explicit representation of water molecules. A solvent layer of 8.0 Å around the molecule was modelled; for the calculation, *periodic boundary conditions* were applied.

The denatured state was simulated by the following two methods: (**1**) the complete chain (except the proline residues) was folded into a all-β-strand conformation, with ϕ and ψ angles of 120.0 degrees for each amino acid, resulting in a long, stretched chain with *maximum solvent accessibility*. (**2**) Only the α-helical and β-strand regions (e.g. residues 7 to 19, 23 to 30; 1 to 4, 32 to 35) were folded into the stretched conformation, whereas the regions of turns and loops were kept fixed relative to the native state. This resulted in a more compact starting conformation.

The simulations described under (**2**) were divided in two different parts: (a) the fixation of the turns was released immediately after the equilibration time of the dynamics; (b) turns were kept fixed for the first 40 ps of the dynamics. Under the latter conditions the chain has already collapsed to a compact structure before the turns are allowed to decompose.

Beginning with the unfolded states described under (**1**), changes in the backbone dihedral angles of residue Glu-23 (the central residue of the molecule) were introduced to examine the influence of small *perturbations* on the resulting structure. An additional torsion of +30° and -60° for the ϕ-angle of residue Glu-23 resulted in three slightly different initial structures. All other angles in the three structures were identical.

RESULTS AND DISCUSSION

The simulations performed during this work show that the final state is reached after about 40 to 50 ps. After this time a compact structure has emerged; the total energies of the structures are very similar to each other and to the energy of the native state. It may be argued that protein folding takes place within seconds instead of picoseconds. In fact, it cannot be ruled out that there may be major rearrangements of the structures on a time scale 10^6 larger than the one used here, even under the same conditions. However, a close inspection of the last 30 picoseconds of the simulations (Fig. 1) suggests that there is a high energy barrier (activation energy) for a significant transition between the final four structures.

A comparison of the dynamic trajectories with explicit representation of water molecules and analogous calculations *in vacuo* (data not shown) revealed that the resulting conformations and trajectories are completely different, as it is expected. *Electrostatic shielding* effects as well as *solvent damping* of the motions of the protein atoms significantly delay the collapse of the chain to a compact globule. Intramolecular interactions compete with molecule-solvent interactions; therefore, the analysis of the simulation gets more complex and much more computer time is needed for a simulation.

It was not the aim of this work to investigate the influence of explicit solvent representation on molecular dynamics simulations[10] but to examine the effect of small structural perturbations on dynamical events under identical conditions; therefore, the limitation to *in vacuo* conditions - whatever arguments exist contrary to those calculations - was justifiable. Further work will include (aside from other improvements in model representation) explicit water representation.

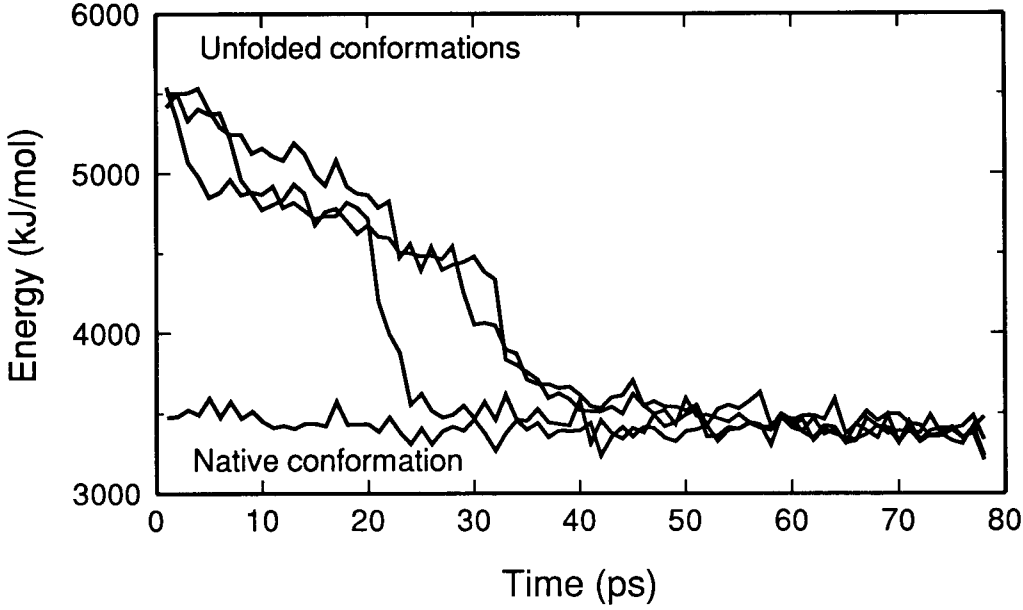

Fig. 1: Total energy (kinetic and potential energy) during 'refolding' simulation of crambin.

Fig. 1 demonstrates that the total energy of the system (including potential and kinetic energy) drops during the simulated folding process. When a compact state is reached, only minor fluctuations of the energy (and the structure, see Fig. 2) take place. The energy in the final state is identical within these fluctuation to the native state of the molecule, e.g. they are energetically indistuingishable. This holds also true when several terms of the energy (*hydrogen bonding energy, van der Waals energy, Coulombic energy, dispersion energy*) are compared, respectively.

The structures, however, are divergent during the simulation; this is indicated in Fig. 2 and 3. Depending on the initial structure, all three simulations result in structures which are far away from the native protein. The starting conformation is about 40 Å rms deviation away from the native (reference) conformation; all structures end up about 8 to 10 Å rms deviation from the native state. This describes the collapse of the stretched chain to a folded, compact globule. Fig. 3 demonstrates that the evolution of the structures during the collapse diverges; the final structures are about 8 to 10 Å rms deviation away from each other.

A closer examination reveals that this is the maximal deviation that compact globular structures of the given size can adopt. Therefore, the three final structures of the simulations and the native state represent four completely different protein conformations.

Surprisingly, the simulated structures contain locally secondary-structure like regions of short α-helices and β-sheets; native proteins show more regular local geometries, but the characteristic hydrogen bonding patterns can be identified unambiguously. Unfortunately they are in different spatial regions in all structures. An analysis of the formation of those secondary structures demonstrate that the decision for the local structures to occur origins between 10 ps and 25 ps; sidechain interactions of spatially neighboring amino acids do have a great influence on the tendency for the structures to occur, although the final structures are stabilized by hydrogen bonds of the protein backbones. The main reason for the distinct structures seem to be the sum of occurences of subtle differences which occur '*by chance*'.

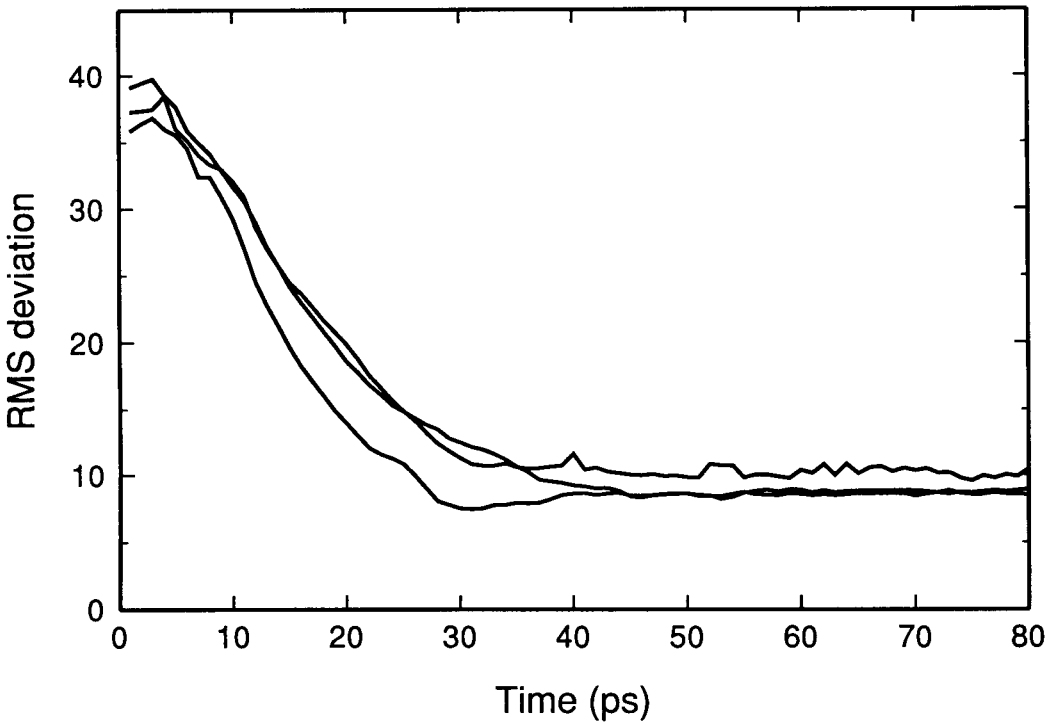

Fig. 2: Root mean square deviation of the three unfolded conformations of crambin to the native state.

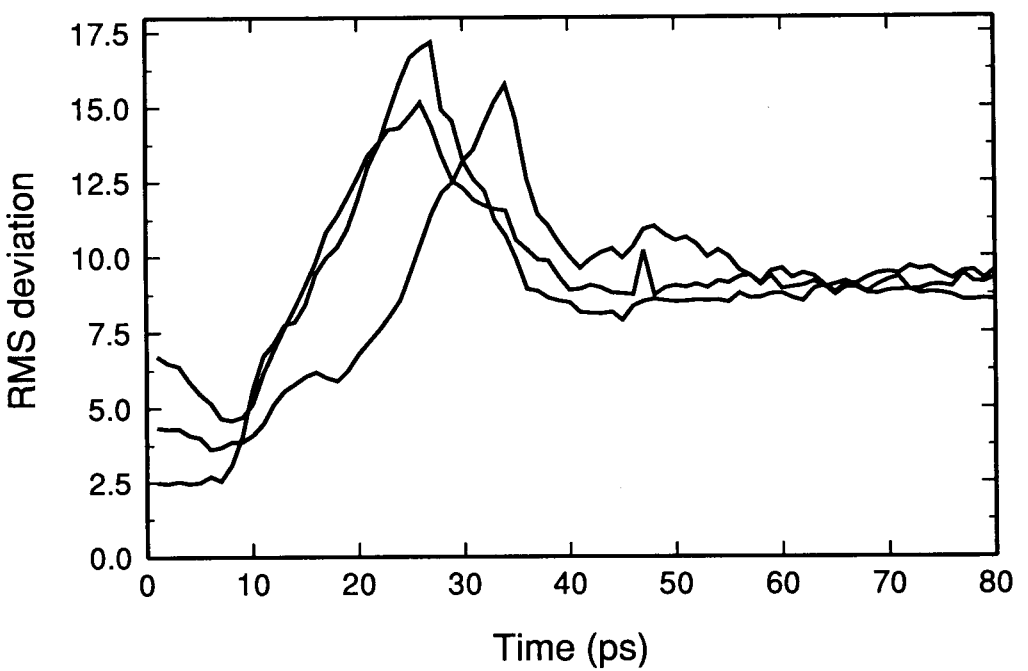

Fig. 3: Root mean square deviation of the three unfolded conformations of crambin relative to each other.

CONCLUSIONS

The divergence of the trajectories of the molecules during the simulations may be considered to be mainly dependend on the selected parameters and the algorithms used, but this seems not to be the case since the results are independent from all guidelines and rules examined so far. The structures observed converge in terms of the calculated enthalpic terms (*potential energy*), e.g. the native state is energetically indistuingishable from the computationally refolded molecules starting from three slightly different states; however, the structures do not converge but diverge to completely different states.

This does in no case proof, of course, that for all starting conditions and for all deterministic algorithms which are possible the results are similarly discouraging. Especially it must be taken into account that the underlying basis of the force fields and the *Newtonian dynamics* which is used are primitive mechanical models and are definitely too simple to be reliable for the description of a complex process such as the simulation of protein folding[11].

On the other hand, there is no imminent reason to believe that our current axiomatic system of mathematics and physics does really provide a solution to the protein folding problem. The existence of a unique relationship between sequence and structure of a protein does not necessarily imply that this relationship can be expressed in our mathematical or physical axiomatic system; this holds true for several multi-body problems of modern science.

Now, what would be the benefit of the analogous formulation of the protein folding problem to previously described systems of deterministic chaos? One achievement is the classification of structure calculation methods which have been published recently. A graphical standard procedure is the description of dynamical processes *via bifurcation graphs*; true bifurcation graphs are characterized by the Feigenbaum constant. Some ideas, e.g. the module method, may be shown to work only since they circumvent the information mass by reducing the folding process to a section of the conformational hyperspace. Others, e.g. the built-up-procedure, starts from a completely different point in the bifurcation graph and thus from another point in hyperspace. Both methods, however, lack a general applicability since they reduce conformational information without providing the required energy.

There are properties describing and characterizing chaotic systems which may also be useful for the judgement of the value of new methods. The *Lyapunov-exponent* characterizes the nature of the 'final state' (attractor) of the chaotic system; if the exponent is negative, the attractor for the trajectories is a fixed point in the phase space of conformations (static model, e.g. crystal structure). If it is zero, then the native conformation would be in equilibrium between several transition states (dynamic model, e.g. equivalent substates of hemoglobin); finally, a positive Lyapunov-exponent would characterize a '*strange attractor*' which has no known equivalent in protein structure theory so far.

Also, the *chaotic dimension* is a valuable information for the description of dynamical processes. This parameter establishes the conformational space for a molecule to fold; it is equivalent to the so-called conformational hyperspace. It would be of great interest to find out correlations between the chaotic dimension and statistical parameters (for example, the size of the protein, or amino acid composition). Even the order of magnitude for this value is still controversal. The *fractal index*[12,13] may be used to characterize the softness of the surface of proteins; chemical reactivity is often related with a rough surface[13]. Also, methods have been developed that relate protein C_α-atoms and self-affine fractal surfaces[15]. The theory of *renormalization*

groups connects physical forces on completely different scales; this is useful for comparing thermodynamical (experimental) results with molecular structure calculations[16].

Bifurcation diagrams are helpful schemes for the description of dynamical trajectories and parameter dependence of processes and may be of value for the description of transition states of the folding or intermediate states which are detected by experimental methods. The folding pathway described *via* bifurcation diagrams are not directly related to kinetical schemes for the refolding of proteins which are gained by experiments; these schemes are based on thermodynamical criteria rather than molecular-structural data and thus describe the properties of an ensemble of molecules, not the fate of a single molecule. Now, bifurcation diagrams do provide a simple possibility for the description of multiple protein folding pathways. It has been observed that the refolding pathway and the occurrence of the observable intermediates (in a thermodynamical sense these intermediates are an ensemble of molecules exhibiting the same behavior of a measurable parameter) may be different when refolding starts from different conditions. Finding out how the Feigenbaum constant can be derived from the experimental values of refolding from different denatured states should provide us with a better understanding of the folding problem.

In a recent work, El Nashie and Kapitaniak found that the distinction between chaotic and strange nonchaotic behavior may be performed by the Lyapunov exponent distribution of symbolic dynamic simulations[17]. The authors stress the similarity of the experimental results from nucleic acids research and chaotic models of solitons in elastic strings.

The current work identifies analogous facts between known highly chaotic systems and the protein folding problem. However, at the current state we cannot rule out that the description in terms of the chaos theory is incomplete, even if there is evidence. The general applicability of the methods described for the molecular dynamics simulation is definitely inadequate. However, to my knowledge there is no contradiction to experimental or computational results in the literature. On the other hand, tools like *fractal indices, chaotic dimensions, bifurcation diagrams, and Lyapunov-exponents* may be of value for protein structure prediction.

If protein folding can be identified unambiguously as a process of deterministic chaos, then this would imply that any *ab initio* method must fail when the native state of a protein is the kinetically accessible minimum but not the thermodynamically possible minimum intrinsically contained in the amino acid chain. In this case, knowledge-based methods like *homology modelling, genetic algorithms, neural network* methods or other pattern recognition techniques are the only way to calculate protein structure from sequence. Therefore, future work should concentrate on those methods rather than the improvement of force field parameters or force field algorithms[18,19].

Acknowledgements - The aid of many colleagues is gratefully acknowledged. Special thanks to Profs. Gustav Obermair, Rainer Rudolph, and mostly Rainer Jaenicke for help and kind support; Drs. Miklós Cserzö, Ina Koch, and Cornelius Frömmel for fruitful discussions. This work was supported by Deutsche Forschungsgemeinschaft grant Ja 78/29-1.

REFERENCES

1 Jaenicke, R., Is There a Code for Protein Folding ? in *Protein Structure and Protein Engineering*, 39. Colloquium - Mosbach 1988 (ed.: Winnacker, E.-L., & Huber, R.). Springer Verlag, Berlin/Heidelberg/New York, pp. 16-36 (1988)

2 Jaenicke, R., Folding and Association of Proteins. *Prog. Biophys. molec. Biol.* **49**, 117-237 (1987).

3 Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., & Thornton, J.M., Knowledge-based Prediction of Protein Structures and the Design of Novel Molecules. *Nature* **326**, 347-352 (1987).
4 Vásquez, M., & Scheraga, H.A., Calculation of Protein Conformation by the Build-up Procedure. Application to Bovine Pancreatic Trypsin Inhibitor Using Limited Dimulaten Nuclear Magnetic Resonance Data. *J. Biomol. Struct. Dynam.* **5**, 705-755 (1988).
5 Skolnick, J., & Kolinski, A., Simulation of the Folding of a Globular Protein. *Science* **250**, 1121-1125 (1990).
6 Lorenz, E., Deterministic Nonperiodic Flow, *J. Atmosph. Sci.* **20**, 130-r141 (1963).
7 Schuster, H.G., *Deterministic Chaos: An Introduction*. VCH Publishers, Deerfield/Weinheim (1984).
8 Crick, F., The Recent Excitement about Neural Networks. *Nature* **337**, 129-132 (1989).
9 Goldberg, D.E., *Algorithms in Searching, Optimization, and Machine Learning*. Addison-Wesley (1989).
10 van Gunsteren, W.F., & Berendsen, H.J.C., Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem. Int. Ed. Engl.* **29**, 992-1023 (1990).
11 Frauenfelder, H., Biomolecules, in *Emerging Syntheses in Science Volume I* (ed.: Pines, D.). Addison-Wesley (1988).
12 Mandelbrot, B., *The Fractal Geometry of Nature*. Freeman, New York (1977).
13 Li, H., Li, Y., & Zhao, H., Fractal Analysis of Protein Chain Conformation. *Int. J. Biol. Macromol.* **12**, p. 6-8 (1990).
14 Bryant, S.H., Islam, S.A., & Weaver, D.L., The Surface Area of Monomeric Proteins: Significance of Power Law Behavior. *Proteins Struct. Funct. Genet.* **6**, 418-423 (1989).
15 Cserzö, M., & Vicsek, T., Self-Affine Fractal Analysis of Protein Structure. *Chaos, Solitons and Fractals* **1**, in press (1991).
16 Fisher, M.E., The Renormalization Group in the Theory of Critical Behavior. *Reviews of Modern Physics* **46**, 579-616 (1974).
17 El Naschie, M.S. & Kapitaniak, T., Soliton Chaos Models for Mechanical and Biological Elastic Chains. *Physics Letters A* **147**, 275-281 (1990)
18 Garnier, J., Protein Structure Prediction. *Biochimie* **72**, 513-524 (1990).
19 Fasman, G.D., *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York (1989).