Low Energy Conformations of a Three-Helix Peptide in an All-Atom Biomolecular Forcefield

T. Herges, A. Schug and W. Wenzel

Forschungszentrum Karlsruhe
Institut fr Nanotechnologie
Postfach 3640, D-76021 Karlsruhe, Germany

ABSTRACT

Using a recently developed all-atom forcefields for biomolecular structure prediction we have analyzed an approximate free-energy surface of the 36 residue headpiece of the villin protein with stochastic optimization methods. With an initial paramterization of the solvent accessible surface area based solvation term we found configurations that were lower in energy than the NMR configuration. We then adjusted the parameters of the solvent model to stabilize the NMR structure using a decoy approach and arrived at a free energy surface that is characterized by a deep folding funnel populated by different three helix structures one of which is very similar to the NMR structure.

1 Introduction

Biomolecular structure prediction remains one of the main outstanding problems of theoretical biophysical chemistry[1]. One of its primary goals is the prediction of the three-dimensional, tertiary structure of proteins on the basis of their amino acid sequence (protein structure prediction — PSP). Experimental methods for protein structure determination are orders of magnitude more involved and more expensive than sequencing techniques. Although their number is steadily growing, the protein database (PDB), presently contains about 13,000 spatially resolved structures[2].

Theoretical methods for PSP may be helpful to close this gap, but accurate theoretical methods that would permit a routine prediction of this structure remain elusive, in particular at the *ab-intio* level. In this approach one simulates a model for the protein, which accounts for the intramolecular interactions of the molecule and its interactions with the environment. Failure to fold may thus stem from two principle sources: the forcefield used to model the protein may be inaccurate or the simulation technique may be inefficient.

In this paper we report on the exploration of the free-energy surface of the 36 amino-acid headpiece (HP36) of the the villin protein (pdb-code: 1VII) with an all-atom forcefield in an attempt to distinguish between these two possiblities. HP36, an autonomously folding peptide, has received much theoretical attention, since it was the subject of one of the most ambitious attempts to simulate the folding process using molecular dynamics[3], which ultimately failed despite an enormous investment of computational resources. In our simulations we used an all-atom protein forcefield (PFF01) [13] with an implicit solvent model. Based on an initial fit of the solvent model parameters we have investigated the free energy surface (FES) of the HP36 and found a two helix-structure, similar to those reported in[3], that was lower in free energy than relaxed NMR structures. The failure to stabilize the latter could thus be attributed to deficiencies of the forcefield rather than to the optimization method. We then recalibrated the forcefield parameters in an attempt to stabilize the three-helix NMR structure. We find that this modified FES has a deep folding funnel that is dominated by three-helix structures, even though the NMR structure is still only a metastable configuration.

2 Methodology

Traditional simulation techniques, in particular molecular dynamics, have great difficulty to access the timescale of protein folding. An obvious starting point for an improved treatment[4] is the elimination of the explicit treatment of the solvent molecules[5], which often consumes the majority of the numerical effort associated with the simulation of the overall system. Upon closer inspection of this approximation, we find that the introduction of an implicit solvent model has far deeper implications on PSP than the obvious reduction of the computational effort resulting from the reduction of the degrees of freedom of the simulation. We note that the overwhelming majority of the entropic contribution to the folding process are solvent contributions, mediated by the hydrophobic and hydrophilic effects of the different amino acid side chains. Incorporating these terms into an implicit solvent model we obtain in conjuction with the internal energy of the protein a good model for the total *free energy* of the system[6]. As indicated above most proteins attain a unique stable native structure. If the protein is in thermodynamic equilibrium with its environment, this structure must therefore correspond to the global minimum of its free energy surface[4]. As is well known from the simulation of many physical systems with complex dynamics, it is possible

to locate the thermodynamically stable state of the system using *stochastic optimization methods* without recourse to its dynamics *orders of magnitude faster* than in an simulation approach[7], [8].

2.1 Biomolecular Forcefield

Over the last decades many classical forcefields [9]–[12] have been developed to investigate numerous phenomena in physical, organic and inorganic chemistry. The difficulties encountered in PSP justify the development of specific forcefields for the following reasons: Their molecular building blocks, i.e. the amino acids, are well defined and limited in number. The chemical complexity associated with the design of a forcefield specific to peptides and proteins is therefore less than that of generic organic substances.

The details of the PFF01 have been described elsewhere[13], here we summarize its main ingredients. The PFF01 forcefield represents all atoms except apolar CH_n individually. CH_n groups are approximated by a single sphere comprising both the carbon and the hydrogen atoms (united atom approach). We have fitted the LJ radii in PFF01 to a subset of 134 proteins of the PDB database. The associated LJ interaction strength was taken from the OPLS forcefield[14]. We note that in simulations with explicit solvent molecules there are LJ interactions between peptide and solvent atoms. This atom-dependent effect has been incorporated into the implicit solvent model. Coulomb interactions are modeled with group-dependent and interaction dependent effective dielectric constants[15]. For the implicit solvent model, the simplest conceivable choice assigns a free energy of solvation proportional to the effective contact area each atom of the protein/peptide has with the solvent. We have subdivided the atom types of the forcefield into suitable subgroups and fitted the resulting model to the available experimental Gly-X-Gly data[5].

3 Results

We have investigated the 36 residue headpiece of the villin protein that was recently simulated with molecular dynamics[3]. The best configuration obtained with about a CPU week on a single PC is shown in Figure 2(b) in comparison with the NMR structure. The fraction of native contacts was similar in both studies. This comparison illustrates the increase in efficiency that can be obtained through the use of stochastic optimization methods, even though both simulations failed to reach the NMR structure. We find however that the structure obtained in our simulation has a lower (free) energy that that of the NMR structure, indicating that this failure is not due to a failure of the optimization strategy, but is attributable to a shortcoming of the forcefield.

This suggests a rational decoy strategy to systematically improve the forcefield the we presently implement.

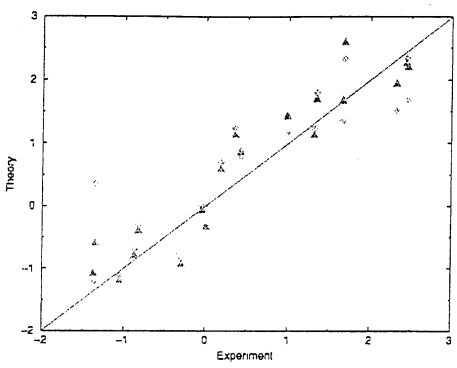

Figure 1: Correlation between the free energies of solvation between experimental data for Gly-X-Gly and two solvent accessible surface area based models (in units of kcal/mol) that differ in the number of atom groups used in the fit. The PPF01 forcefield uses the fit indicated by the triangles with an RMS error of less than 0.5 kcal/mol.

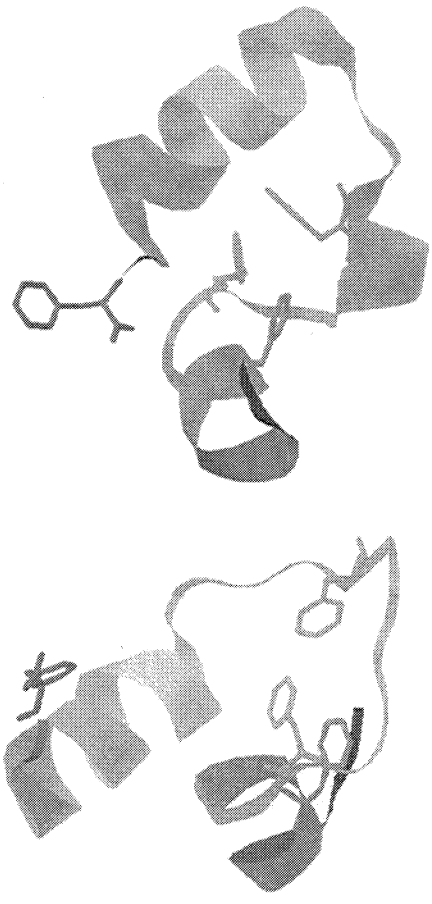

Figure 2: Comparison of the (a) NMR structure and the (b) simulated structure of 1VII.

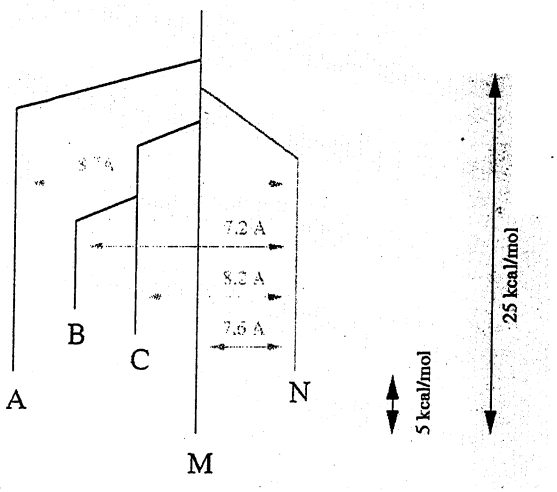

Figure 3: Schematic structure of the low-energy part of the free-energy surface of HP36/1VII in the refined energy model.

We generate a large set of "good" candidates that compete with the NMR structure. As long as one of these decoys has a better energy than the native configuration, the forcefield must be modified to stabilize the native configuration in comparison to all other decoys. When this is achieved we generate new decoys by refolding the peptide, generating new configurations that are either yet again better in energy than the NMR structure or ultimately folding the peptide. In the following we report the preliminary results of this project.

We have created a set of decoys starting either with stretched configurations or the NMR configuration. Some of the latter runs were modified with an additional harmonic constraint that limited the deviation of the simulated structure to the NMR structure to 2-3 A. The adjustable forcefield parameters were the surface free energies that enter the implicit solvent model, which were permitted to vary by 20% around their original values. The rationale behind this approach was that these parameters are relatively uncertain, as they are transferred from small-molecule data to very large systems.

Using this approach we finally arrived at a decoy set containing about 11,000 entries that each had a backbone RMSD of at least 1 A with every other decoy. This decoy set yields an approximative representation of the local minima of the FES of the peptide. The best configuration had a ficticious free-energy of -83.0 kcal/mol (see Fig.4 (M)), the best NMR-like configuration (see Fig.4 (N)) ranked number 3 of 11,000 and had a backbone RMSD of 3.6 A to the NMR configuration.

In order to analyze the FES we classified the configurations into families as a function of energy in order to generate a tree-like structure that has previously been used to analyze complex PES[16]. For all decoys below a given energy we classify two decoys as belonging to the same family if their backbone RMSD is less than 3 A.

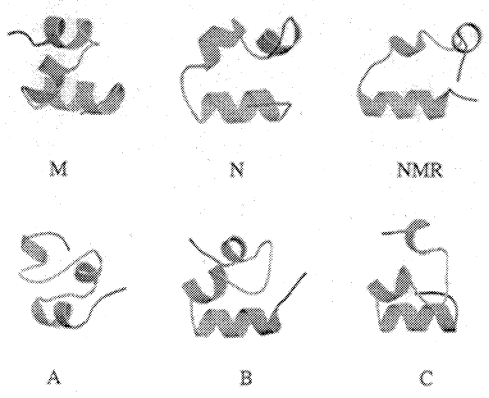

Figure 4: Representatives of the best decoy families of HP36/1VII

We then follow these families by increasing the energy threshold. The resulting graph of the FES is shown in Fig.3. At the bottom we find only the minimal structure found in the simulations, which we label M, with an energy of -83.0 kcal/mol. This configuration is shown in Fig. 4(M). As we increase the energy the set of decoys, to which M belongs grows in number. At about -78 kcal/mol two additional sets of decoys appear, which we label N and A respectively. The set N contains a single configuration that is relatively close to the NMR structure (backbone RMSD 3.6 A) and which we used as a reference point to measure the backbone RMSD to other families. The set A contains a three helix configuration with comparable energy (-78.7 kcal/mol) and a backbone RMSD of 8.7 A to decoy N. As we increase the energy two events can occur: First, new families appear which have higher minimal energies. Secondly, branches of the tree eventually unite as each family grows and family membership is associative. The backbone RMSD between the energetically lowest member of each family to decoy N are indicated by the arrows.

The main results of this analysis can be summarized as follows: There are only very few distinct low-lying minima of the FES of our model. One of those is a good representative of an NMR-like structure, although it is still a metastable state. Upon closer inspection all low-lying branches of the FES correspond to three helix structures (see Table1). The secondary structure content of the minimal structure M is closer to the NMR structure than the NMR decoy.

Overall this picture is consistent with the existence of one very complex folding funnel in the FES. From the standpoint of secondary structure analysis this funnel is characterized as containing only three-helix structures. Within the folding funnel the configuration explores a subspace of the full FES in which helix length and position vary. Surprisingly there is almost no correlation in the RMSD between these structures. We note that only configurations with a deviation of more than 1 A were

Decoy Code	Secondary Structure
NMR	cctHHHHHbbbbbtHHHHttbtHHHHHHHHHHbbcc
N	cctHHHHHHHbbtHHHHHHHHHHHtHHHHHHHHHHcc
M	cctHHHHHHHbbbbHHHHHHcctHHHHHHHHHHccc
A	ccbtHHHHHHHcbtHHHHHHbccbbbbtHHHHtccc
B	cctHHHHHHHHHHtHHHHHHcbbcccbtHHHHHtcc
C	ccbtHHHHHHHHHHtHHHHHbcbbtbcctHHHHbbcc

Table 1: Secondary structure analysis (MOLMOL) of the terminal representatives of the branches of the FES depicted in Fig.3

counted as individual decoys. With backbone RMS deviations in excess of 7 A, the families of low-lying structures differ as much among one another as they differ with a random configuration of the decoy set.

We also note that the energies at which the different branches of the tree unite are much higher than thermally accessible transition states. Ideally the tree of configurations should be constructed using estimates of the transition states between different families[17], [16], but this is numerically prohibitive for the large number of decoy configurations considered here. For the lowest four decoys such searches are presently conducted to evaluate the accuracy of the tree. The present path connecting two families is therefore only an upper bound on the transition state energy. It is very likely that members of different families are bridged by partially unfolded configurations that are lower in energy but do not appear in the decoy database because they are no local minima of the FES.

4 Summary and Conclusions

We have motivated the use of stochastic optimization methods as a technique to predict the structure of complicated biomolecules. To implement this approach, we have developed a biomolecular forcefield, PFF01, that parameterizes the free energy of the underlying system with an implicit representation of the interactions of the biomolecule with the solvent. We have argued that there is a rational, decoy-based strategy to develop a biomolecular forcefield that can be used to predict the structure of short peptide fragments using stochastic optimization techniques such as the stochastic tunneling method. We have illustrated this approach in the folding of short peptide fragments and presented an analysis of the difficulties encountered in the folding of the 36 head residues of 1VII. We have demonstrated that stochastic optimization methods permit an analysis of this peptide at the all-atom level and open a systematic route for the improvement of the forcefield. We note that the search for an optimal forcefield for HP36/1VII, which stabilizes the native configuration as the absolute minimum of the FES, was not exhaustive and is presently still under way.

Acknowledgments: This work was funded by the Deutsche Forschungsgemeinschaft (We 1863/11-1), the BMWF and the Bode foundation.

REFERENCES

[1] D. Baker and A. Sali, Science **294**, 93 (2001).
[2] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, Nucleic Acids Research **28**, 235 (2000), http://www.rcsb.org/pdb.
[3] Y. Duan and P. A. Kollman, Science **23**, 740 (1998).
[4] J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kamierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye, *et al.*, Proc. Natl. Acad. Sci. (USA) **98**, 2329 (2001).
[5] D. Eisenberg and A. McLachlan, Nature **319**, 199 (1986).
[6] M. Daune, *Molecular Biophysics: Structures in Motion* (Oxford Scientific, 1999).
[7] B. A. Berg and T. Neuhaus, Phys. Letters **B267**, 249 (1991).
[8] K. Binder and A. Young, Rev. Mod. Phys. **58**(4), 801 (1986).
[9] W. van Gunsteren and H. Berendsen, *The Groningen Molecular Simulation Manual (GROMOS)*, Tech. Rep., Groningen University (1987).
[10] M. MacKerell, J. Phys. Chem. **B102**, 3586 (1998).
[11] W. L. Jorgensen and N. A. McDonald, J. Mol. Struct. **424**, 145 (1998).
[12] Y. Duan, L. Wang, and P. Kollman, Proc. Nat. Acad. Science USA **95**, 9897 (1995).
[13] T. Herges, H. Merlitz, and W. Wenzel, J. Ass. Lab. Autom. **7**, 98 (2002).
[14] W. L. Jorgensen, Encyclopedia of Computational Chemistry **3** (1998).
[15] F. Avbelj and J. Moult, Biochemistry **34**, 755 (1995).
[16] D. J. Wales, M. A. Miller, and T. R. Walsh, Nature **394**, 758 (1998).
[17] D. J. Wales, J. Chem. Phys. **91**(11), 7002 (1989).