John von Neumann Institute for Computing

Optimization of a Physics-Based United-Residue Force Field (UNRES) for Protein Folding Simulations

A. Liwo, C. Czaplewski, St. Ołdziej, U. Kozłowska, M. Makowski, S. Kalinowski, R. Kaźmierkiewicz, H. Shen, G. Maisuradze, H. A. Scheraga

published in

NIC Symposium 2008,
G. Münster, D. Wolf, M. Kremer (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **39**, ISBN 978-3-9810843-5-1, pp. 63-70, 2008.

http://www.fz-juelich.de/nic-series/volume39

Optimization of a Physics-Based United-Residue Force Field (UNRES) for Protein Folding Simulations

Adam Liwo[1,3], Cezary Czaplewski[1,3], Stanisław Ołdziej[2,3], Urszula Kozłowska[3], Mariusz Makowski[1,3], Sebastian Kalinowski[1], Rajmund Kaźmierkiewicz[2,3], Hujun Shen[3], Gia Maisuradze[3], and Harold A. Scheraga[3]

[1] Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland
E-mail: {*adam, czarek, momo, bes*}*@chem.univ.gda.pl*

[2] Intercollegiate Faculty of Biotechnology, University of Gdańsk, Medical University of Gdańsk
ul. Kładki 24, 80-822 Gdańsk, Poland
E-mail: {*stan, rajmund*}*@biotech.ug.gda.pl*

[3] Baker Laboratory of Chemistry and Chemical Biology, Cornell University
Ithaca, NY 14853-1301, U.S.A.
E-mail: {*uad2, hs322, gm56, has5*}*@cornell.edu*

Understanding the functioning of living cells requires knowledge of structure and long-time dynamics of proteins and other biological macromolecules, which information is not readily available from experiment. The development of distributed computing has opened new avenues for such studies. Further, reduction of the representation of polypeptide chains to the so-called united-residue or coarse-grained representation enables the extension of the time scale of calculations to micro- or even milliseconds. In this report, we describe recent developments of the united-residue (UNRES) force field for large-scale simulations of protein structures and dynamics carried out with the use of the resources at the Supercomputer Centre in Jülich.

1 Introduction

One of the major and still unsolved problems of computational biology is to understand how interatomic forces determine how proteins fold into the three-dimensional structures. The practical aspect of the research on this problem is to design a reliable algorithm for the prediction of the three-dimensional structure of a protein from its amino-acid sequence, which is of utmost importance because experimental methods for determination of protein structures cover only about 10% of new protein sequences. In the case of protein structure prediction, methods that implement direct information from structural data bases (e.g., homology modelling and threading) are, to date, more successful compared to physics-based methods[1]; however only the latter will enable us to extend the application to simulate protein folding and to understand the folding and structure-formation process. The underlying principle of physics-based methods is the *thermodynamic hypothesis* formulated by Anfinsen[2], according to which the ensemble called the "native structure" of a protein constitutes the basin with the lowest free energy under given conditions. Thus, energy-based protein structure prediction is formulated in terms of a search for the basin with the lowest free energy; in a simpler approach the task is defined as searching for the conformation with the lowest potential energy[3], and prediction of the folding pathways can be formulated as a search for the family of minimum-action pathways leading to this basin from the unfolded (denaturated) state. In neither procedure do we want to make use of ancillary data

from protein structural databases. Equally important is to simulate the pathways of protein folding, misfolding (which is the cause of prion diseases, cancer, and amyloid diseases) and large-scale conformational changes which occur during enzymatic catalysis or signal transduction.

United-residue (also termed coarse-grained or mesoscopic) representations of polypeptide chains enable us to carry out large-scale simulations of protein folding, to study protein free-energy landscapes and to carry out physics-based predictions of protein structure[4]. Owing to the considerable reduction of the number of interacting sites and variables, the cost of computations is reduced hundreds or thousand of times compared to an all-atom representation of polypeptide chains in implicit and explicit solvent, respectively; this enables micro- or even millisecond simulations of protein folding to be carried out. For the past several years, we have been developing a physics-based coarse-grained UNRES model of polypeptide chains and the corresponding force field[5–14]. Initially[7], it was designed for physics-based predictions of protein structure through global optimization of an effective potential-energy function of polypeptide chains plus solvent. With this approach, we achieved considerable success in blind-prediction CASP exercises[12]. Recently[15] we implemented a mesoscopic dynamics method to the UNRES force field which enabled us to carry out real-time *ab initio* simulations of protein folding. Subsequently, we implemented[16] the replica-exchange (REMD)[17] and multiplexing-replica exchange (MREMD)[18] extensions of MD, which enabled us to study the thermodynamics of protein folding. However, the new applications required reparameterization of the UNRES force field to reproduce the thermodynamic characteristics of protein folding.

2 Methods

2.1 The UNRES Model of Polypeptide Chains

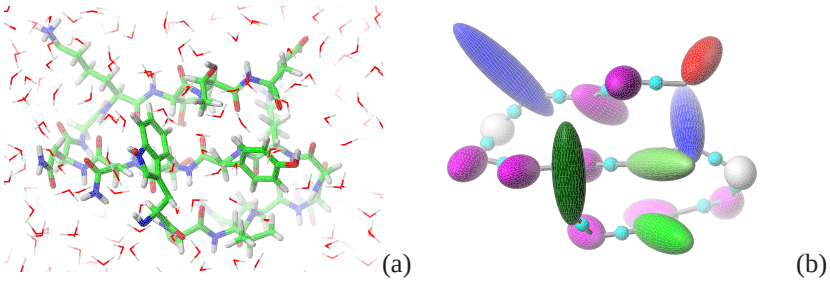

Figure 1. Illustration of the correspondence between the all-atom polypeptide chain in water (a) and its UNRES representation (b). The side chains in part (b) are represented by ellipsoids of revolution and the peptide groups are represented by small spheres in the middle between consecutive α-carbon atoms. The solvent is implicit in the UNRES model.

In the UNRES model[5–14], 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). Each united peptide group is located in the middle between two consecutive

α-carbons. Only these united peptide groups and the united side chains serve as interaction sites, the α-carbons serving only to define the chain geometry, as shown in Figure 1. The $C^{\alpha} \cdots C^{\alpha}$ virtual bond lenghts (i.e., the distances between neighbouring C^{α}'s) are 3.8 Å corresponding to *trans* peptide groups.

The effective energy function is a sum of different terms corresponding to interactions between the SC ($U_{SC_iSC_j}$), SC and p ($U_{SC_ip_j}$), and p ($U_{p_ip_j}$) sites, as well as local terms corresponding to bending of virtual-bond angles θ (U_b), side-chain rotamers (U_{rot}), virtual-bond torsional (U_{tor}) and double-torsional (U_{tord}) terms, virtual-bond-stretching (U_{bond}) terms, correlation terms ($U_{corr}^{(m)}$) pertaining to coupling between backbone-local and backbone-electrostatic interactions[8] (where m denotes the order of correlation), and a term accounting for the energetics of disulfide bonds (U_{SS}). Each of these terms is multiplied by an appropriate weight, w, which must be determined by optimization of the energy function by using training proteins. The energy function is given by equation 1.

$$\begin{aligned} U = & w_{SC} \sum_{i<j} U_{SC_iSC_j} + w_{SCp} \sum_{i \neq j} U_{SC_ip_j} + w_{pp} \sum_{i<j-1} U_{p_ip_j} \\ & + w_{tor} \sum_i U_{tor}(\gamma_i) + w_{tord} \sum_i U_{tord}(\gamma_i, \gamma_{i+1}) \\ & + w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) + \sum_{m=3}^{6} w_{corr}^{(m)} U_{corr}^{(m)} \\ & + w_{bond} \sum_{i=1}^{nbond} U_{bond}(d_i) + w_{SS} \sum_i U_{SS;\, i} \end{aligned} \tag{1}$$

The method of optimizing the force field developed in our laboratory is termed *hierarchical method*[11,14] and aims at obtaining such energy landscapes of selected training proteins that the free energy of each of the training proteins decreases with increasing native likeness. The conformational space is discretized into levels, each of which corresponds to a certain degree of native likeness. In the present study we computed the free energies below, at, and above the folding-transition temperatures and extended the approach by the requirements that the free-energy relations be inverted above the folding-transition temperature. This is illustrated in Figure 2.

Given a set of training proteins, optimization of the force field consists of iterating the cycles, each consisting of (i) simulations with current parameters of the energy function, (ii) computing the free energies at selected temperatures from the simulation data, and (iii) adjusting the parameters of the energy function to achieve the desired relations between the free energies of the sub-ensembles of the training proteins (Figure 2). The procedure is terminated when the required relations between free energies hold after a new simulation with optimized parameters. The primary optimizable parameters were the energy-term weights of equation 1. To generate the decoy sets for optimization at various temperatures simultaneously, we implemented the multiplexing replica-exchange molecular dynamics (MREMD)[17,18] in UNRES[16]. We developed a parallel well-scalable code for the UNRES/MREMD method, which scales 75% up to 4096 processors (Figure 3). To compute free energies and other ensemble-related quantities from simulation data, we implemented the weighted-histogram analysis method (WHAM)[19].

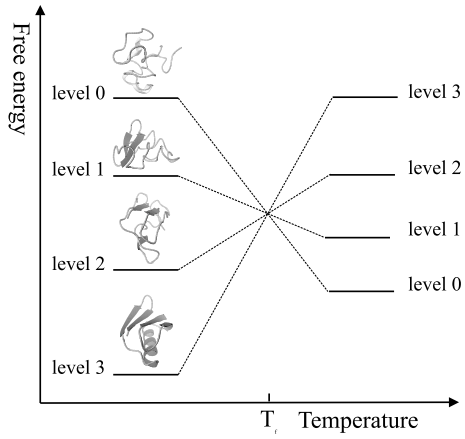

Figure 2. Illustration of ordering of the energy levels, which is the goal of the algorithm for optimizing the potential function[14], using the 1EM7 protein of the IGG family, the structure of which consists of two β-hairpins packed to a middle α-helix. Only one conformation has been selected to represent each of the structural levels. Below the folding-transition temperature (T_f), the non-native level (level 0) has the highest free energy, the conformations with only the native C-terminal β-hairpin forming (level 1) have a lower free energy, next are the conformations in which the middle part of the N-terminal β-strand joins the β-hairpin and the middle α-helix starts to form and, finally, the native-like structures with all structural elements formed have the lowest free energy. Above the folding-transition temperature the free-energy relations are reversed and at the folding-transition temperature the free energies should be approximately equal.

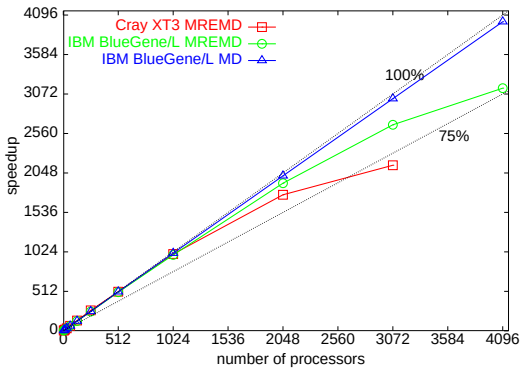

Figure 3. Speedup plots for the MREMD code using IBM Blue Gene/L (green circles) and Cray XT3 (red rectangles). For comparison, data from an ideally scalable series of independent canonical MD runs are shown. The system is the 1SAP protein.

3 Results

Initially[14] we applied the new optimization procedure to three proteins separately identified by the following PDB codes: 1E0L (a 28-residue anti-parallel three-stranded β-sheet), 1GAB (a 47-residue three-α-helix bundle), and 1E0G (a 48-residue $\alpha + \beta$ protein). All three force fields exhibited a heat-capacity peak corresponding to the folding-transition

temperature. The force field optimized on 1GAB was fairly transferable to other α-helical proteins, whose native-like ensembles of structures were located within the five most probable clusters of structures[14].

As the next step, we carried out hierarchical optimization using two training proteins: 1ENH (a three-helix bundle; 56 residues) and the full 37-residue sequence of 1E0L. This choice was motivated by the availability of the experimental temperature dependence of the free energy[20,21] and by the fact that these proteins, although small in size, contain significant regions with undefined secondary structure, which are hard to reproduce in simulations. To generate a converged statistical ensemble of conformations for one protein, typically 20,000,000 MREMD steps with 1024 processors are required, which means 1 rack-week of computation with Blue Gene/L. The total computational effort to optimize the force field on 1E0L and 1ENH was 2 rack-months with Blue Gene/L.

The most probable conformations of the two training proteins calculated with the optimized parameters of the UNRES energy function at room temperature are superposed on the experimental structure on Figure 4. The experimental[20,21] and calculated free-energy gaps vs. temperature are compared in Figure 5, while the calculated heat-capacity and RMSD curves vs. temperature are shown in Figure 6.

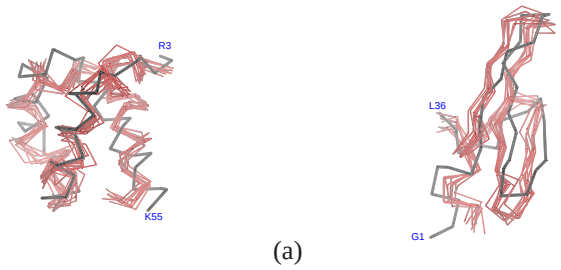

Figure 4. The C^{α} traces of 10 most probable conformations at T=300°K of 1ENH (a) and 1E0L (b) calculated with the UNRES force field optimized on these two proteins (red lines) superposed on the C^{α} traces of the corresponding experimental structures (black lines)[20,21].

With the optimized force field, we carried out MREMD simulations of the mutants of 1E0L studied by Gruebele et al.[20] (these mutations result in a shift of the folding temperature). Subsequently, we calculated the heat-capacity curves of 1E0L mutants and determined their folding temperatures. All mutants studied except for the W30A mutant folded to structures similar to that of the wild-type protein (with ensemble-averaged RMSD at room temperature from 3.5 to 4.5 Å). The calculated folding temperatures are compared with their experimental counterparts in Table 1.

4 Conclusions

The results of our research demonstrated that it is possible to obtain a coarse-grained force field for protein simulations which reproduces thermodynamic properties of wild-type proteins and their mutants, as well as is transferable to proteins outside the training set. Consequently, large-scale simulations of protein structure and dynamics are at hand. Our current

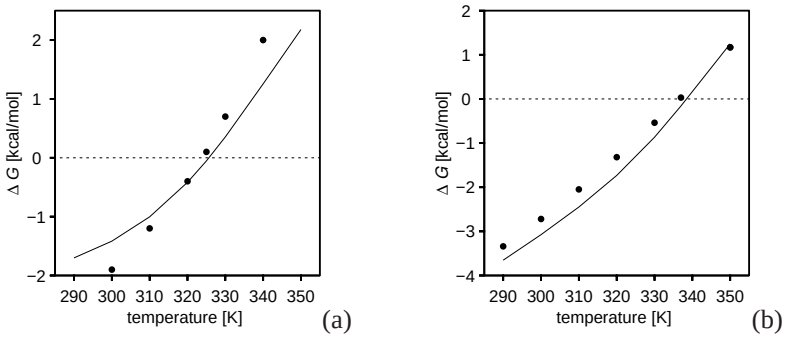

Figure 5. Calculated with optimized force field (lines) and experimentally determined (filled circles) free energy of folding of (a) 1ENH and (b) 1E0L.

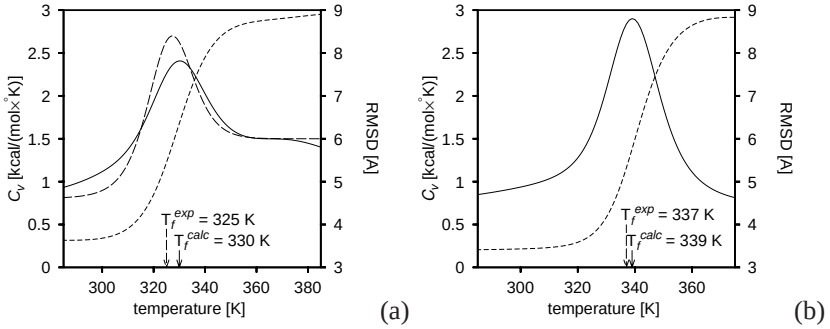

Figure 6. Calculated heat-capacity curves (solid lines) and RMSD curves vs. temperature (short-dashed lines) of (a) 1ENH and (b) 1E0L. The long-dashed curve shown in panel (a) is the experimental heat-capacity curve of 1ENH shifted vertically to match the tail of the calculated heat-capacity curves. The calculated and experimental folding temperatures are also shown. The experimental heat-capacity curve of 1E0L has not been determined.

research is focused on improving the parameterization of UNRES and using the force field to study the kinetics of protein folding, protein aggregation, and large-scale motions.

Acknowledgments

This research was conducted by using the resources of the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Jülich, Germany and was financially supported by grants from the National Institutes of Health (GM-14312), the National Science Foundation (MCB05-41633), the NIH Fogarty International Center (TW7193), and grant DS 8372-4-0138-7 from the Polish Ministry of Science and Higher Education.

Mutant	T_f^{calc}	T_f^{exp}
WT	339	337
W30F	334	339
W30A	–[a]	328
Y11R	342	339
Y19L	319	328
DNY11R[b]	335	339
DNDCY11R[c]	325	328

Table 1. Experimental and calculated, with the optimized force field, folding temperatures of mutants of 1E0L.

[a]This mutant did not fold in MREMD simulations.
[b]Deletion of the 6-residue N-terminal fragment.
[c]Deletion of the N-terminal and the C-terminal fragments.

References

1. J. Moult. Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Phil. Trans. R. Soc. B*, 361:453–458, 2006.
2. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
3. H. A. Scheraga, J. Lee, J. Pillardy, Y.-J. Ye, A. Liwo, and D. R. Ripoll. Surmounting the multiple-minima problem in protein folding. *J. Global Optimization*, 15:235–260, 1999.
4. S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein. Coarse grain models and the computer simulations of soft materials. *J. Phys. Condens. Matter*, 16:R481–R512, 2004.
5. A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.*, 18:849–873, 1997.
6. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Ołdziej, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. *J. Comput. Chem.*, 18:874–887, 1997.
7. A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci., U. S. A.*, 96:5482–5485, 1999.
8. A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.*, 115:2323–2347, 2001.
9. S. Ołdziej, U. Kozłowska, A. Liwo, and H. A. Scheraga. Determination of the potentials of mean force for rotation about $C^\alpha \cdots C^\alpha$ virtual bonds in polypeptides from the *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline. *J. Phys. Chem. A*, 107:8035–8046, 2003.

10. A. Liwo, S. Ołdziej, C. Czaplewski, U. Kozłowska, and H. A. Scheraga. Parameterization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from *ab initio* energy surfaces of model systems. *J. Phys. Chem. B*, 108:9421–9438, 2004.
11. S. Ołdziej, A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape: II. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B*, 108:16934–16949, 2004.
12. S. Ołdziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kaźmierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y.-K. Kang, K. D. Gibson, and H. A. Scheraga. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field – test with CASP5 and CASP6 targets. *Proc. Natl. Acad. Sci. U.S.A.*, 102:7547–7552, 2005.
13. M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H. A. Scheraga. Molecular dynamics with the united-residue (UNRES) model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J. Phys. Chem. B*, 109:13785–13797, 2005.
14. A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, and H.A. Scheraga. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B*, 111:260–285, 2007.
15. A. Liwo, M. Khalili, and H. A. Scheraga. Molecular dynamics with the united-residue (UNRES) model of polypeptide chains; test of the approach on model proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2362–2367, 2005.
16. M. Nanias, C. Czaplewski, and H. A. Scheraga. Replica exchange and multicanonical algorithms with the coarse-grained united-residue (UNRES) force field. *J. Chem. Theor. Comput.*, 2:513–528, 2006.
17. U. H. E. Hansmann, Y. Okamoto, and F. Eisenmenger. Molecular dynamics, Langevin and hybrid Monte Carlo simulations in multicanonical ensemble. *Chem. Phys. Lett.*, 259:321–330, 1996.
18. Y. M. Rhee and V. S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys J.*, 84:775–786, 2003.
19. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations of biomolecules. I. The method. *J. Comput. Chem.*, 13:1011–1021, 1992.
20. H. Nguyen, M. Jäger, A. Moretto, M. Gruebele, and J. W. Kelly. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *Proc. Natl. Acad. Sci. U.S.A.*, 100:3948–3953, 2003.
21. U. Mayor, J. G. Grossman, N. W. Foster, S. M. V. Freund, and A. R. Fersht. The denaturated state of engrailed homeodomain under denaturating and native conditions. *J. Mol. Biol.*, 333:977–991, 2003.