



Evolution of Experimental and Theoretical Determinations of Protein Structure and Protein Folding Pathways

H. A. Scheraga, A. Liwo, C. Czaplewski, S. Ołdziej

published in

*From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,*
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. 36, ISBN 978-3-9810843-2-0, pp. 45-54, 2007.

© 2007 by John von Neumann Institute for Computing
Permission to make digital or hard copies of portions of this work for
personal or classroom use is granted provided that the copies are not
made or distributed for profit or commercial advantage and that copies
bear this notice and the full citation on the first page. To copy otherwise
requires prior specific permission by the publisher mentioned above.

<http://www.fz-juelich.de/nic-series/volume36>

Evolution of Experimental and Theoretical Determinations of Protein Structure and Protein Folding Pathways

Harold A. Scheraga, Adam Liwo, Cezary Czaplewski, and Stanisław Ołdziej

Baker Laboratory of Chemistry and Chemical Biology, Cornell University,
Ithaca, NY 14853-1301, U.S.A.
E-mail: has5@cornell.edu

Physical chemical studies of hydrogen bonding and hydrophobic interactions, and experimental studies of the structure and folding pathways of bovine pancreatic ribonuclease A motivated the development of a theoretical approach to compute protein structure and protein-folding pathways.

1 Introduction

This article traces the development of our experimental and theoretical efforts to gain an understanding of the underlying physics that controls the progression from a newly-synthesized polypeptide chain to the three-dimensional structure of a native biologically-active fibrous or globular protein. Our earliest involvement with this problem was concerned with the influence of hydrogen bonds and hydrophobic interactions on protein structure and reactivity. This work led to our efforts to determine protein structure and folding pathways, first by experimental methods, and subsequently by theoretical methods.

2 Internal Bonding in Proteins

Internal hydrogen bonds influence the observed pKs of ionizable groups¹ and even the reactivity of covalent bonds², e.g., peptide bonds. Figure 1 provides an example of a hydrogen bond between a tyrosyl donor and a glutamate acceptor. The observed pKs of

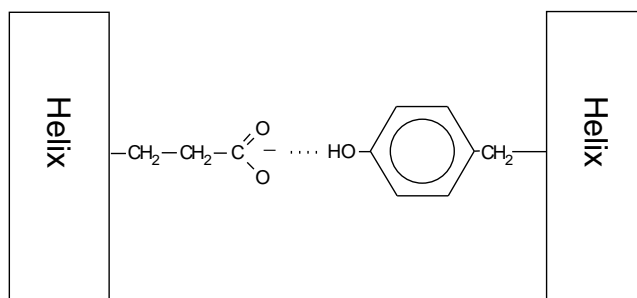


Figure 1. Example of a tyrosyl \cdots glutamate hydrogen bond.

these groups are modified by the free energy to form (or break) such a hydrogen bond. Therefore, in comparison with a non-hydrogen-bonded model compound, the observed pKs of such hydrogen-bonded tyrosyl and glutamate groups will be raised or lowered, respectively. Consequently, such departures of the pKs from those of model compounds are diagnostic for the presence of such hydrogen bonds.

Hydrophobic interactions can provide a nonpolar environment which will also influence the pKs of nearby ionizable groups. A theory³ for the thermodynamics of hydrophobic interactions, based on the structures of liquid water and of aqueous hydrocarbon solutions was presented in 1962, and upgraded⁴ in 2004. By themselves, hydrogen bonds in proteins in water are not very strong because of the necessity to shed water in order to form a hydrogen bond between the polar groups of a protein. However, as illustrated in Figure 2, the presence of nearby nonpolar groups can provide hydrophobic interactions⁵ with the nonpolar parts of residues such as lysine and glutamic acid and also restrict the internal rotational freedom of the ionizable side chains. In addition, nonpolar groups can restrict the access of water to the polar parts of ionizable side chains. Thus, the cooperativity of nonpolar groups and hydrogen bonding of ionizable side chains can strengthen the hydrogen bonds.

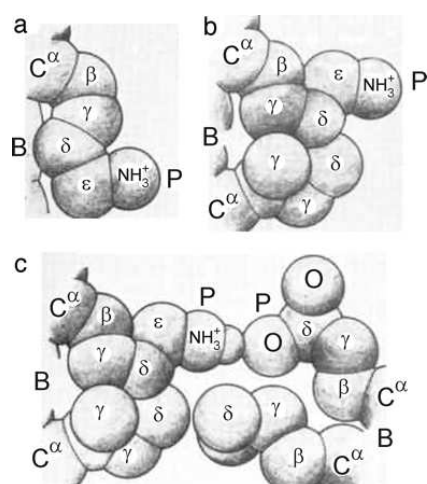


Figure 2. Illustration of various hydrophobic interactions of a polar side chain with its surroundings. B refers to the backbone, and P to the polar head.

3 Location of Hydrogen Bonds in Proteins

Before the advent of X-ray crystallography, NMR, and recombinant DNA methods to determine protein structure, experimental studies to locate hydrogen bonds between ionizable groups, as indicated by the dotted lines in Figure 3, provided distance constraints on the folding of a protein backbone. Such studies, carried out on the 124-residue protein bovine pancreatic ribonuclease A (RNase A), showed that 3 of its 6 tyrosyl residues had

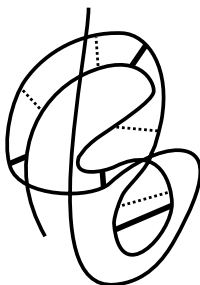


Figure 3. Schematic representation of a protein. Solid and dotted lines represent disulfide bonds and non-covalent interactions, respectively.

abnormally-high⁶ pKs, and that 3 of its 11 carboxyl groups had abnormally-low⁷ pKs. Further, the UV absorption spectrum of tyrosine was perturbed⁸ at low pH where carboxyl groups ionize. This evidence suggested the existence of three tyrosyl \cdots carboxylate hydrogen bonds, and subsequent physical and biochemical experiments⁷ identified one pairing, namely

Tyr25 \cdots Asp14
Tyr92 \cdots Asp38
Tyr97 \cdots Asp83

out of the 19,800 possible ways to pair 3 of 6 Tyr and 3 of 11 carboxyl groups of RNase A. The identification of these 3 interactions was subsequently verified by the X-ray structure. These three non-covalent interactions (dotted lines in Figure 3) and the four disulfide bonds (solid-line crosslinks in Figure 3) provide 7 distance constraints on the folding of the backbone. However, 7 distance constraints are not sufficient to provide an accurate description of the backbone of a 124-residue protein such as RNase A. To determine the backbone structure, as is now done by NMR, many more distance constraints would be required. In fact, it is possible to specify the number of distance constraints required⁹ in order to determine the structure within any desired RMSD from the native structure.

4 Initial Considerations of a Theoretical Approach to Structure Simulation

On the other hand, even 7 known distances could serve as restraints on a potential energy function to compute the native structure of a protein. This provided the motivation to develop¹⁰ a theoretical approach to compute protein structure, first by making use of distance restraints and, subsequently, to rely on a physics-based potential function without the need to incorporate distance restraints. At about the same time, Anfinsen¹¹ identified spontaneous protein folding, and introduced the thermodynamic hypothesis for a theoretical approach, and we expanded our interest from determining structure to also determining folding pathways (first by experiment, and later by theory).

5 Experimental Studies of Oxidative Folding of RNase A

Our experimental study of the oxidative folding of RNase A led to the mechanism shown in Figure 4. Figure 4(a) shows that a pre-equilibrium exists between the unfolded forms

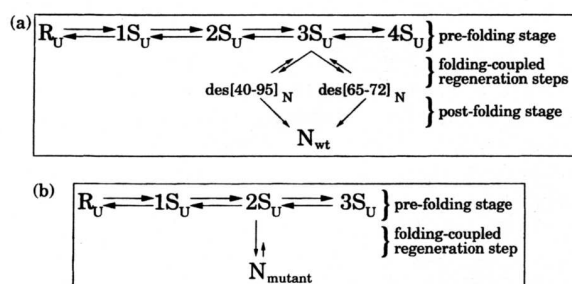


Figure 4. (a) Oxidative folding of wild-type RNase A. (b) Oxidative folding of two three-disulfide mutants of RNase A (C40A/C95A and C65S/C72S).

of reduced RNase A, R , and various ensembles of disulfide-bonded intermediates¹². The rate-determining step¹³ is the reshuffling of the three-disulfide ensemble, $3S$, by SH/SS interchange, to form two main intermediates, $des [40-95]_N$ and $des [65-72]_N$ which each contain three native disulfide bonds but lack the 40-95 and 64-72 disulfide bonds, respectively. These two native intermediates^{14,15} rapidly form the native structure of the wild-type protein. As shown in Figure 4(b), two very minor pathways exist^{16,17} in which the $2S$ ensemble undergoes oxidation to $des [40-95]_N$ and $des [65-72]_N$, which could be detected only with the aid of mutants which lacked the 40-95 and 65-72 disulfide bonds, respectively.

The overall scheme for the oxidative folding of RNase A is shown in Figure 5. Of the 28 possible $1S$ species, 40% have the native 65-72 disulfide bond, and 10% have the non-native 58-65 disulfide bond, and the remaining 26 species accumulate only to the extent of $<10\%$ each in folding of the whole protein¹⁸. The 65-72 disulfide bond persists increasingly in the remainder of the pathway¹⁹ to form $des [40-95]_N$. Interestingly, the same 40:10 ratio that is found in the protein is also found when a fragment of reduced RNase A from Cys 58 to Cys 72 is oxidized^{20,21}. This result is attributed to preferential native-forming interactions²² and not to entropic effects in the 65-72 loop, since both possible loops (58-65 and 65-72) have the same size; it is this kind of physics that is revealed by such experimental studies, and by our concomitantly developed molecular mechanics approach.

6 All-atom Determination of Protein Structure and Folding Pathways

Progressing from our initial work¹⁰ in 1965, with a hard-sphere potential, we developed an all-atom ECEPP (Emperical Conformational Energy Program for Peptides) force field²³,

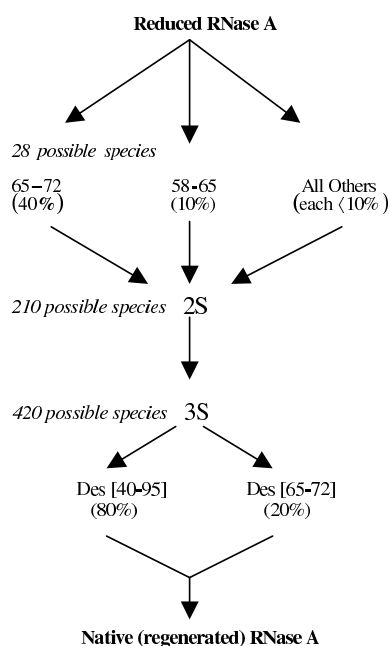


Figure 5. Overall scheme of oxidative folding of RNase A

and improved it in subsequent years, the latest version of which²⁴ appeared in 2006. The ECEPP force field and various procedures that we developed for global optimization of the potential energy were described in a recent review²⁵. The largest protein whose structure we have computed with the all-atom force field is the 46-residue three-helix bundle, protein A²⁶, and the 36-residue villin headpiece²⁷.

Our initial attempt²⁸ to compute an all-atom folding pathway made use of the stochastic difference equation method of Elber²⁹. The pathways were computed from each of a large ensemble of unfolded states of protein A to the final folded state, and then averaged. It was found that the C-terminal helix folded first, followed by the N-terminal helix, and then the middle helix. Various folding pathways have been proposed for protein A, and it has been found³⁰ that environmental factors and different components in the various force fields used may account for the reported differences.

7 A Hierarchical Approach to Protein Structure and Folding Pathway Prediction

In order to compute protein structures larger than those of protein A, we have developed a hierarchical procedure which initially makes use of a united-residue (UNRES) model of a polypeptide chain³¹⁻³⁷ together with a conformational space annealing (CSA) procedure³⁸ to search the UNRES conformational space to find the *region* in which the global minimum might lie. Then the lowest-energy structures are converted from the UNRES representation

to an all-atom one^{39,40} whose ECEPP energy (including hydration) can then be globally optimized.

The UNRES model consists of virtual-bond chains for the backbone and side chains, with the less-important degrees of freedom (rotation of the peptide groups around their virtual C^α-C^α bonds, internal rotations about side-chain bonds, etc.) averaged out. The force centers are the positions of the averaged-out peptide groups and the ends of the virtual bonds at the center of gravity of the side chains. The UNRES energy consists of interactions between these force centers, the energies to vary the positions and rotational states of side chains, the energies to vary the angle between successive backbone virtual bonds, the torsional angles around the backbone virtual bonds, and double torsions around two neighboring virtual bonds, and multi-body interactions. The CSA procedure is essentially a genetic algorithm in which a finite set of widely- dispersed UNRES minima are forced to coalesce to the region of the global minimum.

Performance in successive blind tests from CASP3 to CASP7 has provided sufficient confidence to encourage us to develop^{41,42} and apply⁴³⁻⁴⁵ a molecular dynamics treatment based on UNRES. Our recent work with this molecular dynamics approach is being discussed at this workshop by A. Liwo⁴⁶.

8 Conclusions

The evolution of the development of our experimental and theoretical approaches to gain an understanding of the fundamental physics that controls protein structure and folding pathways has been traced. It is elaborated upon in the accompanying article by Liwo et al⁴⁶. Current work is focused on the use of the molecular dynamics approach with UNRES, and the improvement of this methodology including introduction of entropic effects³⁷.

Acknowledgments

This research was supported by grants from NIH (GM-14312, TW-7193) and NSF (MCB05-41633), and was carried out with resources of (a) our 820-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (c) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Juelich, Germany, (d) the National Center for Supercomputing Applications System at the University of Illinois at Urbana-Champaign, (e) the cluster at the Department of Computer Science, Cornell University, (f) the resources of the Center for Computation and Technology at Louisiana State University, which is supported by funding from the Louisiana legislature's Information Technology Initiative, (g) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdansk, (h) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdansk, and (i) the Cornell Theory Center which receives funding from Cornell University, New York State, Federal agencies, and corporate partners.

References

1. M. Laskowski, Jr. and H.A. Scheraga *Thermodynamic considerations of protein reactions. I. Modified reactivity of polar groups*, J. Am. Chem. Soc. **76**, 6305–6319, 1954.
2. M. Laskowski, Jr. and H.A. Scheraga *Thermodynamic considerations of protein reactions. II. Modified reactivity of primary valence bonds*, J. Am. Chem. Soc. **78**, 5793–5798, 1956.
3. G. Némethy and H.A. Scheraga *The structure of water and hydrophobic bonding in proteins. III. The thermodynamic properties of hydrophobic bonds in proteins*, J. Phys. Chem. **66**, 1773–1789, 1962.
4. J.H. Griffith and H.A. Scheraga *Statistical thermodynamics of aqueous solutions. I. Water structure, solutions with non-polar solutes, and hydrophobic interactions*, J. Molec. Str. **682**, 97–111, 2004.
5. G. Némethy, I.Z. Steinberg, and H.A. Scheraga *The influence of water structure and of hydrophobic interactions on the strength of side-chain hydrogen bonds in proteins*, Biopolymers **1**, 43–69, 1963.
6. C. Tansford, J.D. Hauenstein, and D.G. Rands *Phenolic hydroxyl ionization in proteins. II. Ribonuclease*, J. Am. Chem. Soc. **77**, 6409–6413, 1958.
7. H.A. Scheraga *Structural studies of pancreatic ribonuclease*, Fed. Proc. **26**, 1380–1387, 1967.
8. H.A. Scheraga *Tyrosyl-carboxylate ion hydrogen bonding in ribonuclease*, Biochim. Biophys. Acta **23**, 196–197, 1957.
9. H. Wako and H.A. Scheraga *On the use of distance constraints to fold a protein*, Macromolecules **14**, 961–969, 1981.
10. G. Némethy and H.A. Scheraga *Theoretical determination of sterically allowed conformations of a polypeptide chain by a computer method*, Biopolymers **3**, 155–184, 1965.
11. C.B. Anfinsen *Principles that govern folding of protein chain*, Science **181**, 223–230, 1973.
12. D.M. Rothwarf, Y.-J. Li, and H.A. Scheraga *Regeneration of bovine pancreatic ribonuclease A. Identification of two natively like three- disulfide intermediates involved in separate pathways*, Biochemistry **37**, 3760–3766, 1998.
13. D.M. Rothwarf, Y.-J. Li, and H.A. Scheraga *Regeneration of bovine pancreatic ribonuclease A. Detailed kinetic analysis of two independent folding pathways*, Biochemistry **37**, 3767–3776, 1998.
14. S. Shimotakahara, C.B. Rios, J.H. Laity, D.E. Zimmerman, H.A. Scheraga, and G.T. Montelione *NMR structural analysis of an analog of an intermediate formed in the rate-determining step of one pathway in the oxidative folding of bovine pancreatic ribonuclease A: Automated analysis of ¹H, ¹³C, and ¹⁵N resonance assignments for wild-type and [C65S, C72S] mutant forms*, Biochemistry **36**, 6915–6929, 1997.
15. J.H. Laity, C.C. Lester, S. Shimotakahara, D.E. Zimmerman, G.T. Montelione, and H. A. Scheraga *Structural characterization of an analog of the major rate-determining disulfide folding intermediate of bovine pancreatic ribonuclease A*, Biochemistry **36**, 12683–12699, 1997.

16. M. Iwaoka, D. Juminaga, and H.A. Scheraga *Regeneration of three- disulfide mutants of bovine pancreatic ribonuclease A missing the 65-72 disulfide bond: Characterization of a minor folding pathway of ribonuclease A and kinetic roles of Cys65 and Cys72*, *Biochemistry* **37**, 4490–4501 , 1998.
17. X. Xu and H. A. Scheraga *Kinetic folding pathway of a three-disulfide mutant of bovine pancreatic ribonuclease A missing the [40-95] disulfide bond*, *Biochemistry* **37**, 7561–7571 , 1998.
18. X. Xu, D.M. Rothwarf, and H.A. Scheraga *Nonrandom distribution of the one-disulfide intermediates in the regeneration of ribonuclease A*, *Biochemistry* **35**, 6406–6417 , 1996.
19. M.J. Volles, X. Xu, and H.A. Scheraga *Distribution of disulfide bonds in the two-disulfide intermediates in the regeneration of bovine pancreatic ribonuclease A*, *Biochemistry* **38**, 7284–7293 , 1999.
20. P.J. Milburn and H.A. Scheraga *Local interactions favor the native 8-residue disulfide loop in the oxidation of a fragment corresponding to the sequence Ser-50-Met-79 derived from bovine pancreatic ribonuclease A*, *J. Protein Chem.* **7**, 377–398 , 1988.
21. K.H. Altmann and H.A. Scheraga *Local structure in ribonuclease A. Effect of amino acid substitutions on the preferential formation of the native disulfide loop in synthetic peptides corresponding to residues Cys58-Cys72 of bovine pancreatic ribonuclease A*, *J. Am. Chem. Soc.* **112**, 4926–4931 , 1990.
22. R.P. Carty, M.R. Pincus, and H.A. Scheraga *Interactions that favor the native over the non-native disulfide bond among residues 58-72 in the oxidative folding of bovine pancreatic ribonuclease A*, *Biochemistry* **41**, 14815–14819 , 2002.
23. F.A. Momany, R.F. McGuire, A.W. Burgess, and H.A. Scheraga *Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids*, *J. Phys. Chem.* **79**, 2361–2381 , 1975.
24. Y.A. Arnautova, A. Jagielska, and H.A. Scheraga *A new force field (ECEPP-05) for peptides, proteins and organic molecules*, *J. Phys. Chem. B.* **110**, 5025–5044 , 2006.
25. H.A. Scheraga, A. Liwo, S. Ołdziej, C Czaplewski, J. Pillardy, D.R. Ripoll, J.A. Vila, R. Kazmierkiewicz, J.A. Saunders, Y.A. Arnautova, A. Jagielski, M. Chinchio, and M. Nania *The protein folding problem: Global optimization of force fields*, *Frontiers in Bioscience* **9**, 3296–3323, 2004.
26. J.A. Vila, D.R. Ripoll, and H.A. Scheraga *Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14812–14816 , 2003.
27. D.R. Ripoll, J.A. Vila, and H.A. Scheraga *Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH*, *J. Mol. Biol.* **339**, 915–925 , 2004.
28. A. Ghosh, R. Elber, and H.A. Scheraga *An atomically detailed study of the folding pathways of protein A with the stochastic difference equation*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 10394–10398 , 2002.
29. R. Elber, A. Ghosh, and A. Cardenas *Long Time Dynamics of Complex Systems*, *Acc. Chem. Res.* **35**, 396–403 , 2002.

30. A. Jagielska and H.A. Scheraga *Influence of temperature, friction, and random forces on folding of the B-domain of Staphylococcal Protein A: All-atom molecular dynamics in implicit solvent*, J. Comput. Chem. **28**, 1068–1082, 2007.
31. 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.
32. 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 short-range interactions and determination of weights of energy terms by Z-score optimization*, J. Comput. Chem. **18**, 874–887, 1997.
33. A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Ołdziej, R.J. Wawak, S. Rackovsky, M.R. Pincus, and H.A. Scheraga *A united-residue force field for off-lattice protein-structure simulations; III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials*, J. Comput. Chem. **19**, 259–276, 1998.
34. 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.
35. A. Liwo, P. Arlukowicz, C. Czaplewski, S. Ołdziej, J. Pillardy, and H.A. Scheraga *A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field*, Proc. Natl. Acad. Sci. U.S.A. **99**, 1937–1942, 2002.
36. S. Ołdziej, J. Lagiewka, A. Liwo, C. Czaplewski, M. Chinchio, M. Nancias, and H.A. Scheraga *Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 3. Use of many proteins in optimization*, J. Phys. Chem. B **108**, 16950–16959, 2004.
37. 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.
38. J. Lee, H.A. Scheraga, and S. Rackovsky *New optimization method for conformational energy calculations on polypeptides: Conformational space annealing*, J. Comput. Chem. **18**, 1222–1232, 1997.
39. R. Kazmierkiewicz, A. Liwo, and H.A. Scheraga *Energy-based reconstruction of a protein backbone from its α -carbon trace by a Monte-Carlo method*, J. Comput. Chem. **23**, 715–723, 2002.
40. R. Kazmierkiewicz, A. Liwo, and H.A. Scheraga *Addition of side chains to a known backbone with defined side-chain centroids*, Biophys. Chem. **100**, 261–280, 2003. Erratum: Biophys. Chem., **106**, 91 (2003).
41. M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H.A. Scheraga *Molecular dynamics with the united-residue 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.

42. M. Khalili, A. Liwo, A. Jagielska, and H.A. Scheraga *Molecular dynamics with the united-residue model of polypeptide chains. II. Langevin and Berendsen-bath dynamics and tests on model α -helical systems*, J. Phys. Chem. B. **109**, 13798–13810, 2005.
43. A. Liwo, M. Khalili, and H.A. Scheraga *Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains*, Proc. Natl. Acad. Sci. U.S.A. **102**, 2362–2367, 2005.
44. M. Khalili, A. Liwo, and H.A. Scheraga *Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains*, J. Mol. Biol. **355**, 536–547, 2006.
45. A. V. Rojas, A. Liwo, and H.A. Scheraga *Molecular dynamics with the united-residue force field. Ab initio folding simulations of multichain proteins*, J. Phys. Chem. B. **111**, 293–309, 2007.
46. A. Liwo, C. Czaplewski, S. Ołdziej, M. Chinchio, A.V. Rojas, M. Khalili, M. Makowski, S. Kalinowski, U. Kozłowska, R.K. Murarka, and H.A. Scheraga *Mesoscopic dynamics with the UNRES force field. A tool for studying the kinetics and thermodynamics of protein folding*, in “From Computational Biophysics to Systems Biology 2007”, in press.