In Silico Biology 6 (2006) 15–22
IOS Press

Gauss-Function-Based Model of Hydrophobicity Density in Proteins

Leszek Konieczny[a], Michal Brylinski[b] and Irena Roterman[c,*]

[a]*Institute of Biochemistry, Collegium Medicum, Jagiellonian University, Kopernika 7, 31-034 Cracow, Poland*
E-mail: mbkoniec@cyf-kr.edu.pl
[b]*Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Kopernika 17, 31-501 Cracow, Poland, and Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Cracow, Poland*
E-mail: brylinsk@chemia.uj.edu.pl
[c]*Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Kopernika 17, 31-501 Cracow, Poland, and Faculty of Physics, Jagiellonian University, Reymonta 4, 30-060 Krakow, Poland*
E-mail: myroterm@cyf-kr.edu.pl

Edited by E. Wingender; received 29 September 2005; revised 14 November 2005; accepted 26 November 2005; published 23 January 2006

ABSTRACT: The model adopting the three-dimensional Gauss function to express the hydrophobicity distribution in proteins is presented in this paper. The tendency to create the hydrophobic center during protein folding is expressed in form of an external force field of the form of three-dimensional Gauss function which directs the folding polypeptide to locate the hydrophobic residues in a central part of the molecule and hydrophilic ones exposed toward the molecular surface. The decrease of the differences between hydrophobicity distribution as it appears at each step of the folding simulation and the expected hydrophobicity distribution (three-dimensional Gauss function) is the convergence criterion together with traditional non-bonding interaction optimization. The model was applied to fold the hypothetical membrane protein (target protein in CASP6) TA0354_69_121 from *Thermoplasma acidophilum*.

KEYWORDS: Protein folding, hydrophobicity, hydrophobic collapse, Gauss function, CASP, protein structure prediction

INTRODUCTION

Protein folding is the process in which the starting structure reaches its final structural form (native) through few intermediates [1–4]. The commonly accepted opinion is that the starting (early-stage) steps in folding is backbone conformation dependent [5–7]. One of the intermediate steps is assumed to be hydrophobic collapse-dependent [8–10]. The model to create the early-stage folding structure of a polypeptide has been described elsewhere [11,12]. The geometry-based approach expressing the backbone-dependent conformation of polypeptide chain was discussed in [13] and additionally verified by information entropy analysis [11]. The early-stage folding model resulted in the sequence-to-structure and structure-to-sequence contingency table analysis [14]. The probability values for each fragment (tetrapeptide) to represent a particular structural form allowed for the introduction of a Structure Predictability Index (SPI) [15]. The verification of predicted structures of BPTI [16], ribonuclease [11],

*Corresponding author.

Electronic publication can be found in *In Silico Biol.* **6**, 0002 <http://www.bioinfo.de/isb/2006/06/0002/>, 23 January 2006.

lysozyme [17] and hemoglobin [18] seems to be promising. The early-stage folding (*in silico*) model together with the hydrophobic collapse model is assumed to represent the complete folding process (*in silico*), although the final step of folding process is still missing. It will be presented in a close future.

According to general knowledge of protein characteristics, the participation of hydrophobic interaction in the protein folding process is rather significant [5,19–21]. This is why the model presented in this paper is focused on the structure of hydrophobic core creation during protein *in silico* folding. The 'oil-drop' model introduced by Kauzman [22] visualized well the idea of the tendency to concentrate the hydrophobic residues in the central part of protein molecule with hydrophilic amino acids localized rather on the surface of the molecule. The way in which the amino acid sequence partitions a protein into its inside and outside has been described [23–28].

The model presented here introduces the external hydrophobic force field of the form of *fuzzy-oil-drop* assumed to represent the characteristic hydrophobicity distribution as it is expected in the final structural form of a protein, with the highest density at the center and lowest hydrophobicity density on the surface of the oil-drop. The three-dimensional Gauss function is assumed to represent the *fuzzy-oil-drop* form and to satisfy the expectations concerning the hydrophobicity distribution in a protein molecule. The external hydrophobic force field in the form of a three-dimensional Gauss function is treated as template. Approaching the hydrophobicity density in a protein molecule to this template is expected as a way to reach the structural form observed in the native structure of the protein. The step-wise calculation procedure as applied in this work is presented in Materials and methods.

MATERIALS AND METHODS

Data

Hypothetical membrane protein TA0354_69_121 from *Thermoplasma acidophilum* was selected to prove the reliability of the introduced model. This protein was used as a target molecule (T0215) in CASP6 competition (www.predictioncenter.org). Our group submitted only the early-stage structural form of this protein to verify the reliability of the early-stage structure in blind prediction. The early-stage folding (*in silico*) of this protein was applied to the hydrophobicity-driven process of late-stage protein folding (*in silico*). The simulation of the hydrophobic-collapse-determined intermediate step in the protein folding process adapted to the same molecule was performed to verify the reliability of the presented model.

The simulation of the hydrophobic-collapse-determined intermediate step in the folding process was performed to verify the reliability of the model. In this paper, the crystal form of the protein will be called the "target" molecule, while the structural form predicted will be called the "model".

Fuzzy-oil-drop model – expected hydrophobicity density distribution

The three dimensional Gauss function was applied to represent the distribution of hydrophobicity in a *fuzzy-oil-drop*. The parameters of Gauss function: (1) mean value – (in our case the coordinate x, y, z values) was assumed to be represented by (0,0,0) point in the coordinate system, what means that maximum value of Gauss function is localized in the center of the *fuzzy-oil-drop*; (2) standard deviation – expressing the size of drop (individually for each x, y, z dimension). The value of standard deviation describing the early-stage structural form (starting structure for hydrophobic collapse) was assumed to be polypeptide chain length-dependent of this structural form of polypeptide chain. The standard

deviation parameter is assumed to decrease in a step-wise procedure causing decrease of size of drop and simultaneously the size of molecule. The Gauss function has been standardized at each step of the procedure. The procedure stops, when the size of drop reaches the size of protein molecule of particular number of amino acids in polypeptide chain. The *fuzzy-oil-drop* was represented by grid of appropriate size. Each grid point was described by expected value of hydrophobicity. The expected hydrophobicity value $\tilde{H}e_j$ for a j-th point was calculated as follows:

$$\tilde{H}e_j = \frac{1}{\tilde{H}e_{\mathrm{sum}}} \exp\left(\frac{-(x_j-\bar{x})^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_j-\bar{y})^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_j-\bar{z})^2}{2\sigma_z^2}\right) \quad (1)$$

where σ_x, σ_y, σ_z denote standard deviation and the point $(\bar{x}, \bar{y}, \bar{z})$ possesses the highest hydrophobicity value and has a fixed position in the center of the box (0,0,0) during the simulation. $\tilde{H}e_{\mathrm{sum}}$ is the total hydrophobicity for all analyzed grid points. Each j-th grid point (x_j, y_j, z_j) is characterized by the $\tilde{H}e_j$ value, which represents the idealized density of hydrophobicity in *fuzzy-oil-drop*.

Real, observed hydrophobicity density distribution

The same grid points were described by real, observed hydrophobicity density distribution originating from the folding molecule. For the j-th point the value of hydrophobic density resulted as inter-residual hydrophobic interaction (each amino acid was represented by an effective atom localized in the geometrical center of its side chain). $\tilde{H}o_j$ was calculated according to the Levitt function [29] as follows:

$$\tilde{H}o_j = \frac{1}{\tilde{H}o_{\mathrm{sum}}} \sum_{i=1}^{N} \tilde{H}_i^r \begin{cases} \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right] & \text{for } r_{ij} \leqslant c \\ \text{otherwise} \quad 0 \end{cases} \quad (2)$$

where N is the total number of residues in the protein under consideration, $\tilde{H}_i^r$ denotes the hydrophobicity for the i-th residue according to normalized scale of hydrophobicity for amino acids, r_{ij} denotes the separation of the j-th grid point and the effective atom of the i-th residue. c denotes the hydrophobic cutoff and has the fixed value 9.0Å following the original paper [29]. It means that only residues with $r_{ij} \leqslant c$ influence the j-th grid point. $\tilde{H}o_{\mathrm{sum}}$ is the sum of observed hydrophobicity for all analyzed grid points. Using $1/\tilde{H}o_{\mathrm{sum}}$ as a normalizing coefficient allows the observed hydrophobicity to be compared with the expected hydrophobicity described previously. The hydrophobicity value attached to each grid point expressed the real hydrophobicity density (as determined by folding protein molecule).

Folding procedure

The input file contains the coordinates of atoms for backbone and effective atoms (for side chains of amino acids) of polypeptide chain in its early-stage folding form. The folding procedure was carried out in the following manner:

1. geometrical operation – initial orientation of molecule:
   (a) geometrical center of the molecule is localized in a (0,0,0) point in coordinate system,
   (b) the longest Cα–Cα. distance found in protein molecule determines X-axis,
   (c) the longest distance between Cα–Cα atoms projection on *YZ* plane determines the Y-axis,
   (d) X-axis is oriented according to the standard coordinate system to guarantee the orthogonal order,

(e) all longest distances calculated according to points 1b, 1c and 1d are used to determine the size of molecule and calculate the value of standard deviation for each one-dimensional representation,

2. grid system creation:

   (a) the values of longest distances calculated in 1b, 1c and 1d were enlarged by 9Å,
   (b) the step for grid size is calculated,
   (c) the system of grid points is created,

3. hydrophobicity density:

   (a) each grid point represents the expected hydrophobicity value (fuzzy-oil-drop) calculated according the three-dimensional Gauss function for parameters (standard deviations) as shown in Eq. (1),
   (b) each grid point represents also the observed hydrophobicity value which expresses the real hydrophobicity being the results of inter-residual interactions present in molecule under consideration Eq. (2),
   (c) both scales were standardized (sum of hydrophobicity values for all grid points is equal to 1.0 for both: observed and expected hydrophobicity values),

4. optimization procedure – each step of oil-drop size lowering is followed by structural changes. The structural changes lowering hydrophobicity differences $\Delta\tilde{H}_{\mathrm{tot}}$ were accepted. $\Delta\tilde{H}_{\mathrm{tot}}$ parameter can be interpreted as the difference between expected $\tilde{H}e$ and observed $\tilde{H}o$ hydrophobicity over all grid points j:

$$\Delta\tilde{H}_{\mathrm{tot}} = \sum_{j=1}^{P} (\tilde{H}e_j - \tilde{H}o_j)^2 \tag{3}$$

   where P denotes the total number of grid points at particular step of folding simulation. Each hydrophobicity-based optimization step was followed by energy-based (non-bonding interaction) optimization step. Each of them contained 5000 steps of Rosenbrock's optimization [30]. The energy optimization was introduced to avoid possible overlaps (or high energy orientations).

   (a) hydrophobicity-based optimization:

      – the sum of hydrophobicity differences $\Delta\tilde{H}_{\mathrm{tot}}$ all over grid points (the square) was calculated – the minimization of which is the convergence criterion,
      – the structural transformation (according to degrees of freedom) was performed,
      – the convergence of hydrophobicity density was estimated,

   (b) energy-based (non-bonding interaction) optimization using ECEPP/3 force field [31–35]:

      – structural transformations eliminating possible overlaps were accepted.

The structural changes at each step of the procedure are expressed according to degrees of freedom (Phi, Psi and lambda dihedral angles changes). The iteration has been repeated as long as the size of *fuzzy-oil-drop* reaches the usually observed size of protein molecule containing particular number of amino acids (complete PDB data set was analyzed in respect of molecular size as number of amino acids dependent).

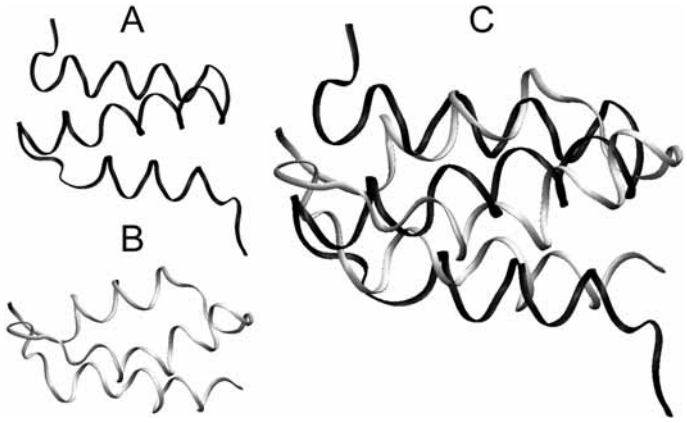

Fig. 1. Ribbon representations of target (A, black) and model (B, grey) structural forms of TA0354_69_121. Superimposed structures are shown in C.

Comparison of crystal structure and predicted structure

The comparison of predicted and real structure of the protein can be performed using two criteria: (1) the relation between predicted and observed protein structure – usually expressed by RMSD factor calculated for Cα atoms and (2) the relation between observed and idealized *fuzzy-oil-drop* hydrophobicity distribution. The first criterion assess the prediction correctness. The second criterion reveals the area of high and low accordance between two objects. The localization of areas with high discrepancy may reveal probable function-related areas in protein molecule.

RESULTS

Early-stage structural form

The early-stage structural form of TA0354_69_121 was predicted from its amino acid sequence according to sequence-to-structure contingency table for tetrapeptides [14] and SPI [15]. The size of the molecule was $57.290 \times 46.182 \times 36.689$Å.

Visual analysis of 3D models

Figure 1 visualizes the structural similarities of native and predicted structural forms of TA0354_69_121. The RMSD-Cα value calculated for the model vs. its crystal structure was found to be 5.656Å.

Non-bonding interactions

Nonbonding interactions present in the target (crystal) and model (predicted) structures are shown in form of contact maps in Fig. 2. Interactions that stabilize the fold are between residues that are well separated along the sequence and therefore away from the diagonal of the plot, where an interaction was defined as occuring when two Cα atoms were within 12Å of one another. Similar patterns of non-bonding interactions were found for target (native) and model (predicted) structures. The percentage of correctly predicted non-bonding contacts was found to be 56.57%.

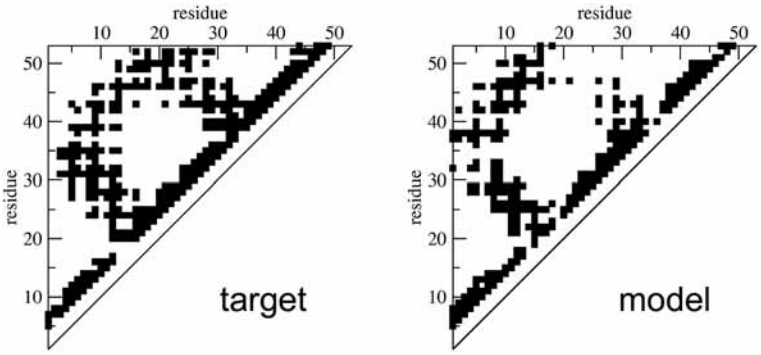

Fig. 2. Non-bonding contact maps for TA0354_69_121.

Biological function localization

Figure 3 shows the distribution of differences between expected and observed hydrophobicity distribution $\Delta\tilde{H}$ along the polypeptide chain. The peaks (Fig. 3A) represent the discrepancy between the expected regular hydrophobicity distribution and the observed one for the protein under consideration. Those fragments are probably function-related parts of a protein (currently the function of TA0354_69_121 is unknown). High accordance of the $\Delta\tilde{H}$ profile found for target and model indicates that the predicted structure (model) reproduces structure and most of the important biological features of the native folded structure (target).

CONCLUSIONS AND DISCUSSION

The idea of hydrophobic collapse as the driving force in the protein folding process is commonly considered to be responsible for tertiary structure determination. The heuristic model of folding simulation driven by the external force field in the form of idealised fuzzy-oil-drop directing the hydrophobic residues to the central part of molecule appeared to lead the molecule toward the native-like structural form. The model of *fuzzy-oil-drop* seems to work well for the *in silico* simulation of hydrophobic collapse.

Few single-domain soluble proteins have been folded using the method presented here (lysozyme [36], ribonuclease [37], BPTI [38]). Model was also used to recognized the active site in proteins [39]. Moreover, the function $1 - \Delta\tilde{H}_{\mathrm{tot}}$ can be applied to large, trans-membrane molecules to push the hydrophobic residues to be exposed toward the membrane.

The presented molecule is all of helical form. The interaction between helices and their mutual orientation is, according to the presented procedure, determined exclusively by non-bonding interaction in the energy-driven step of the optimization procedure and by the hydrophobic interaction with the target *fuzzy-oil-drop* external force field in the hydrophobicity-oriented step. The described molecule was a target in CASP6. Our model for early-stage structure creation applied to 77 targets (the total number of submitted models was 163) revealed this molecule to be our best approach. This is why the late stage model was applied to the same target molecule and treated as the continuation of *blind* prediction. The model is planned to be applied in CASP7 (summer 2006) which will allow the verification of the model in a wide spectrum of different molecules.

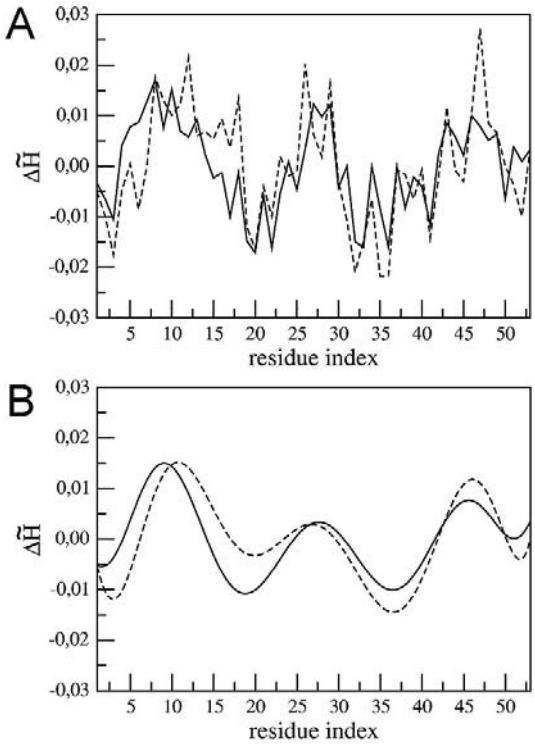

Fig. 3. Profile plots presenting the calculated difference between expected and observed hydrophobicity density distribution $\Delta\tilde{H}$ as a function of the amino acid sequence (A) and its polynomial approximation (B). Solid and dashed lines denote target (native) and model (predicted) structure of TA0354_69_121, respectively.

ACKNOWLEDGEMENTS

Many thanks to Prof. Marek Pawlikowski (Faculty of Chemistry, Jagiellonian University) for fruitful discussions. This research was supported by the Polish State Committee for Scientific Research (KBN) grant 3 T11F 003 28 and Collegium Medicum grants 501/P/133/L and WL/222/P/L.

REFERENCES

[1] Abkevich, V. I., Gutin, A. M. and Shakhnovich, E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. Biochemistry **33**, 10026-10036.
[2] Dill, K. A. and Chan, H. S. (1997). From Levinthal to pathways to funnels. Nat. Struct. Biol. **4**, 10-19.
[3] Roder, H. and Colon, W. (1997). Kinetic role of early intermediates in protein folding. Curr. Opin. Struct. Biol. **7**, 15-28.
[4] Creighton, T. E. (1979). Intermediates in the refolding of reduced ribonuclease A. J. Mol. Biol. **129**, 411-431.
[5] Baldwin, R. L. (2002). Making a network of hydrophobic clusters. Science **295**, 1657-1658.
[6] Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. Philos. Trans. R. Soc. Lond., B, Biol. Sci. **356**, 133-145.
[7] Hayward, S. (2001). Peptide-plane flipping in proteins. Protein Sci. **10**, 2219-2227.
[8] Gutin, A. M., Abkevich, V. I. and Shakhnovich, E. I. (1995). Is burst hydrophobic collapse necessary for protein folding? Biochemistry **34**, 3066-3076.
[9] Qi, P. X., Sosnick, T. R. and Englander, S. W. (1998). The burst phase in ribonuclease A folding and solvent dependence of the unfolded state. Nat. Struct. Biol. **5**, 882-884.
[10] Sosnick, T. R., Mayne, L. and Englander, S. W. (1996). Molecular collapse: the rate-limiting step in two-state cytochrome c folding. Proteins **24**, 413-426.

[11] Jurkowski, W., Brylinski, M., Konieczny, L., Wiiniowski, Z. and Roterman, I. (2004). Conformational subspace in simulation of early-stage protein folding. Proteins **55**, 115-127.
[12] Roterman, I. (1995). Modelling the optimal simulation path in the peptide chain folding-studies based on geometry of alanine heptapeptide. J. Theor. Biol. **177**, 283-288.
[13] Roterman, I. (1995). The geometrical analysis of peptide backbone structure and its local deformations. Biochimie **77**, 204-216.
[14] Brylinski, M., Konieczny, L., Czerwonko, P., Jurkowski, W. and Roterman, I. (2005). Early-stage folding in proteins (*in silico*) sequence-to-structure relation. J. Biomed. Biotechnol. **2**, 65-79.
[15] Brylinski, M., Konieczny, L. and Roterman, I. (2004). SPI – Structure predictability index for protein sequences. In Silico Biol. **5**, 0022.
[16] Brylinski, M., Jurkowski, W., Konieczny, L. and Roterman, I. (2004). Limited conformational space for early-stage protein folding simulation. Bioinformatics **20**, 199-205.
[17] Jurkowski, W., Brylinski, M., Konieczny, L. and Roterman, I. (2004). Lysozyme folded *in silico* according to the limited conformational sub-space. J. Biomol. Struct. Dyn. **22**, 149-158.
[18] Brylinski, M., Jurkowski, W., Konieczny, L. and Roterman, I. (2004). Limitation of conformational space for proteins – early stage folding simulation of human α and β hemoglobin chains. TASK Quarterly **8**, 413-422.
[19] Dill, K. A. (1990). Dominant forces in protein folding. Biochemistry **29**, 7133-7155.
[20] Finney, J. L., Bowron, D. T., Daniel, R. M., Timmins, P. A. and Roberts, M. A. (2003). Molecular and mesoscale structures in hydrophobically driven aqueous solutions. Biophys. Chem. **105**, 391-409.
[21] Pace, C. N., Shirley, B. A., McNutt, M. and Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. FASEB J. **10**, 75-83.
[22] Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. Adv. Protein Chem. **14**, 1-63.
[23] Klapper, M. H. (1971). On the nature of the protein interior. Biochim. Biophys. Acta **229**, 557-566.
[24] Meirovitch, H. and Scheraga, H. A. (1980). Empirical studies of hydrophobicity. 2. Distribution of the hydrophobic, hydrophilic, neutral, and ambivalent amino acids in the interior and exterior layers of native proteins. Macromolecules **13**, 1406-1414.
[25] Meirovitch, H., Rackovsky, S. and Scheraga, H. A. (1980). Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids. Macromolecules **13**, 1398-1405.
[26] Meirovitch, H. and Scheraga, H. A. (1981). Empirical studies of hydrophobicity. 3. Radial distribution of clusters of hydrophobic and hydrophilic amino acids. Macromolecules **14**, 340-345.
[27] Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. Science **229**, 834-838.
[28] Rose, G. D. and Roy, S. (1980). Hydrophobic basis of packing in globular proteins. Proc. Natl. Acad. Sci. USA **77**, 4643-4647.
[29] Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. **104**, 59-107.
[30] Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. Comput. J. **3**, 175-184.
[31] Momany, F. A., McGuire, R. F., Burgess, A. W. and Scheraga, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial pharges, non-bonded interactions, hydrogen bond interactions and intrinsic torsional potentials for naturally ocurring amino acids. J. Phys. Chem. **79**, 2361-2381.
[32] Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S. and Scheraga, H. A. (1992). Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J. Phys. Chem. **96**, 6472-6484.
[33] Dunfield, L. G., Burgess, A. W. and Scheraga, H. A. (1978). Energy parameters in polypeptides. 8. Empirical potential energy algorithm for the conformational analysis of large molecules. J. Phys. Chem. **82**, 2609-2616.
[34] Yan, J. F., Momany, F. A., Hoffmann, R. and Scheraga, H. A. (1970). Energy parameters in polypeptides. II. Semiempirical molecular orbital calculations for model peptides. J. Phys. Chem. **74**, 420-433.
[35] McGuire, R. F., Momany, F. A. and Scheraga, H. A. (1972). Energy parameters in polypeptides. V. An empirical hydrogen bond potential function based on molecular orbital calculations. J. Phys. Chem. **76**, 375-93.
[36] Brylinski, M., Konieczny, L. and Roterman, I. (2006) Fuzzy-oil-drop hydrophobic force field - a model to represent late-stage folding (*in silico*) of lysozyme. J Biomol Struct Dyn **23**, 519-28.
[37] Brylinski, M., Konieczny, L. and Roterman, I. (2006) Hydrophobic collapse (*in silico*) protein folding. Comput Biol Chem, in press.
[38] Brylinski, M., Konieczny, L. and Roterman, I. (2006) Hydrophobic collapse in late-stage folding (*in silico*) of bovine pancreatic trypsin inhibitor. Biochimie, in press.
[39] Brylinski, M., Konieczny, L. and Roterman, I. (2006) Ligation site in proteins recognized *in silico*. Bioinformation 1(4), 127-129.