Transworld Research Network
37/661 (2), Fort P.O., Trivandrum-695 023, Kerala, India


Structure-Function Relation in Proteins, 2009: 79-103 ISBN: 978-81-7895-409-7
Editor: Irena Roterman

# 4 Late stage folding intermediate *in silico* model

**Irena Roterman[1], Leszek Konieczny[2] and Michal Brylinski[1]**
[1]Department of Bioinformatics and Telemedicine – Collegium Medicum – Jagiellonian University, Lazarza 16, 31-530 Krakow, Poland
[2]Institute of Medical Biochemistry – Collegium Medicum – Jagiellonian University, Kopernika 7, 31-034 Krakow, Poland

## Abstract

*The model based on the multi-step character of protein folding process is described in this chapter. The Early Stage (ES) step of protein folding process presented formerly assumed the backbone to be solely responsible for early structural forms. The Late Stage (LS) assumes the hydrophobic interaction as the main driving force for native structure generation. The hydrophobicity is expressed in "in silico" model as external force field of "fuzzy-oil-drop" character generating the environment for folding polypeptide. The three-dimensional Gauss function was taken to represent the hydrophobicity distribution (the original interpretation of the Gauss*

Correspondence/Reprint request: Dr. Irena Roterman, Department of Bioinformatics and Telemedicine – Collegium Medicum – Jagiellonian University, Lazarza 16, 31-530 Krakow, Poland.
E-mail: myroterm@cyf-kr.edu.pl

*function values is interpreted as probability density) concentrating the high hydrophobicity in the center of "fuzzy-oil-drop" with asymptotic zero hydrophobicity density on the surface of the oil-drop with the hydrophobicity decrease accordant to Gauss function. The minimization of the differences between idealized hydrophobicity distribution (Gauss function) and empirical hydrophobicity distribution (dependent on the hydrophobic residues localization in the space) is the basis for optimization procedure. This criterion results as element directing the folding process toward the concentration of hydrophobic residues in a center of the protein molecule with the simultaneous exposure of hydrophilic residues on the protein surface.*

# Introduction

The hydrophobic center in protein molecule has been recognized by Kauzman as the main element responsible for structural stability [1]. It is generally accepted that globular proteins consist of hydrophobic core and the hydrophilic exterior [2-7]. The positioning of particular amino acid as coded by the amino acid sequence has been described [8-11]. The significance of the hydrophobicity distribution along the polypeptide chain increased together with the recognition of the important role of chaperones [12-18]. The dominant role of hydrophobicity has been also recognized experimentally as the dominant driving force for protein folding process [19-22]. The relation between hydrophobicity distribution and packing density as well as the hydrophobicity distribution in 3-D protein structures appeared the main criteria for the evaluation of structures generated ab initio [23,24].

The tertiary structure is understood as the complex form of interaction keeping the protein molecule stable with hydrophobic interaction as the main one. The 'oil-drop' model has been introduced by Kauzman to express the specificity of protein molecules in respect to hydrophobicity and hydrophilicity relation. The modification of this model to the form of "*fuzzy-oil-drop*" is presented in this chapter. Linking all discussed aspects as well as well experimental observations enabled the construction of the external force field expressing the hydrophobicity distribution in form of three-dimensional Gauss function to generate the environment for folding polypeptide. The presence of external force field of this kind directs the hydrophobic residues toward the center of protein molecule with the simultaneous exposure of hydrophilic residues on the surface of the molecule.

# The multi-step character of protein folding process

The experimental observations suggest that the protein folding is the few step process although the number of intermediates is not defined and probably depends on the protein specificity.

The scheme of the process can be presented as follows:

$$U => I_1 => I_2 => \ldots\ldots\ldots => I_k => N$$

The U denotes the unfolded form, I – intermediate and N – native structure of the protein.

The model described in this chapter assumes two intermediates which can be presented as follows:

$$U => ES => LS => N$$

The first intermediate called ES – early stage intermediate has been presented in detailed form formerly in [25]. The ES was generated assuming that solely backbone is responsible for early structural forms of polypeptide. The relation between two geometric parameters: V-angle (dihedral angle between two sequential peptide bond planes) influences the R – radius of curvature of the polypeptide structural form. According to this model all structural forms are of helical character differing by the size of radius of curvature. The β-structural form can be described as the helix of infinitely large radius of curvature with V-angle equal 180 degs. Using this notation, the α-helix is described by the commonly known size of radius of curvature and V-angle equal to zero degs.

The value of V-angle appeared to determine the R – radius of curvature. To avoid the large values of R the logarithmic scale has been introduced. The ln(R) dependency on V-angle appeared to be of second degree polynomial. The part of Ramachandran conformational space, the structures of which satisfy this relation appeared in form of ellipse path linking the structurally significant areas on Ramachandran map revealing the simplest path of structural transformations (probably during the structural changes).

The ellipse path was assumed to represent the limited conformational sub-space. This conformational sub-space appeared to balance the amount of information carried by amino acid sequence and amount of information necessary to define the structural form belonging to ellipse path – the limited conformational sub-space.

All details related to ES model have been given in [26-33].

The structural form of ES intermediate is treated as the starting structure for the simulation of LS – late stage intermediate generation in protein folding process.

## The "*fuzzy-oil-drop*" model for LS intermediate

According to the discussion presented in Introduction, the LS intermediate is generated mostly by the interaction between side chains

(not taken into account in ES intermediate model). Additionally to the traditional interaction expressing the nonbonding interaction (electrostatic, vdW and torsional potential) the hydrophobic interaction has been taken into consideration as the driving force for this step of folding process (according to Kauzman "oil drop" model [1]).

The hydrophobicity is expressed in the form of the presence of external force field of hydrophobic character in the model under consideration [34].

This external force field of hydrophobic character is expressed by the three-dimensional Gauss function. The value of Gauss function traditionally interpreted as probability is assumed to represent the hydrophobicity density. The highest value (in the center of ellipsoid) expresses the highest hydrophobicity decreasing according to Gauss shape reaching the zero value in the distance depending form according to this function. The size of the "*fuzzy-oil-drop*" expressed by the $\sigma_x$ $\sigma_y$ $\sigma_z$ which can take different values for each direction (according to coordinate system). The position of geometric center of the molecule is expressed by $\bar{x}, \bar{y}, \bar{z}$ values.

The three-dimensional Gauss function:

$$G(x,y,x;\bar{x},\bar{y},\bar{z},\sigma_x,\sigma_y,\sigma_z) = K \exp\left(\frac{-(x_j-\bar{x})^2}{2\sigma_x^2}\right)\exp\left(\frac{-(y_j-\bar{y})^2}{2\sigma_y^2}\right)\exp\left(\frac{-(z_j-\bar{z})^2}{2\sigma_z^2}\right)$$

can be applied to describe the external force field of hydrophobic character on the following conditions:

1. the protein molecule (whatever is the structural form of protein – ES in particular) shall be oriented in coordinate system with its geometric center in origin of the coordinate system making the values of $\bar{x}, \bar{y}, \bar{z}$ equal to zero
2. the side chains shall be simplified to one point presentation called as "effective atoms"
3. the molecule shall be oriented to put two residues of the longest distance between them along x-axis
4. the next rotation around y-axis shall be performed to put the two residues of longest distance between them (calculated for points representing the projection on the Y,Z plane of effective atoms).
5. for the orientation as described above the parameters determining the size of ellipsoid expressed by the highest value along x-axis, y-axis and z-axis. These distances increased by 9 Å (cutoff distance for hydrophobic interaction) in each direction. These longest distances (increased by 9 Å

in each direction) divided by 3 produces the values of $\sigma_x$ $\sigma_y$ $\sigma_z$ (according to three-sigma law)

6. for the orientation as describe above the mean values present in Gauss function are equal to zero making the Gauss simplified as follows:

$$\widetilde{Ht}_j = \frac{1}{\widetilde{Ht}_{sum}} \exp\left(\frac{-(x_j)^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_j)^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_j)^2}{2\sigma_z^2}\right)$$

Where: $\widetilde{Ht}_j$ expresses the hydrophobicity in the *j-th* point in space $(x,y,z)$, $\sigma_x$, $\sigma_y$, $\sigma_z$ express the size of *fuzzy-oil-drop*, $\frac{1}{\widetilde{Ht}_{sum}}$ is the normalization coefficient – division by the total hydrophobicity (sum all over the grid points) makes the total hydrophobocity represented by particular "*fuzzy-oil-drop*" normalized to the value of 1.

The folding protein molecule is assumed to approach the idealized hydrophobicity distribution concentrating the hydrophobicity in the central part of molecule with the exposure of hydrophilic residues on the surface of the "*fuzzy-oil-drop*".

The approaching is achieved comparing the idealized hydrophobicity as described above with the observed hydrophobicity distribution depending on the localization of hydrophobic residues, which may be localized in different positions during the folding process.

The grid system with the constant step is generated for the box of $(-3\sigma_x ; +3\sigma_x)$, $(-3\sigma_y ; +3\sigma_y)$, $(-3\sigma_z ; +3\sigma_z)$. The value of hydrophobicity according to idealized hydrophobicity distribution (value of Gauss function) and the value of hydrophobic interaction of each grid point (its own hydrophobicity is equal to zero) collecting the interaction with all side chains (effective atoms) according to following empirical function introduced by M. Levitt [35]:

$$\widetilde{Ho}_j = \frac{1}{\widetilde{Ho}_{sum}} \sum_{i=1}^{N} H_i^r \begin{cases} \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right] & \text{dla } r_{ij} \le c \\ 0 \text{ dla } r_{ij} > c \end{cases}$$

Where $\widetilde{Ho}_j$ denotes the observed hydrophobicity in the *j-th* point in space (in particular the same as in theoretical distribution described in eq. 1.), $r_{ij}$ represents the distance between the *j-th* point and *i-th* which represents the

localization of residue (effective atom carrying the hydrophobicity $H_i^r$) which is specific for particular residue (any hydrophobicity scale available in literature can be applied to this function), the coefficient $\frac{1}{\widetilde{Ho}_{sum}}$ (the total value of hydrophobicity in the denominator) makes the values of $\widetilde{Ho}_j$ normalized to the value 1.

Both - the idealized as well as empirical hydrophobicity density is normalized to the value of 1. It makes possible the comparison of expected and observed hydrophobicity distribution and calculation of the difference between theoretical (idealized) hydrophobicity distribution and empirical one in the point *i-th* :

$$\Delta\widetilde{H}_i = \widetilde{Ht}_i - \widetilde{Ho}_i$$

The minimization of the total difference between expected and observed hydrophobicity is the object for optimization procedure according to the following eq:

$$\Delta\widetilde{H}_{tot} = \sum_{j=1}^{P}\left(\widetilde{Ht}_j - \widetilde{Ho}_j\right)^2$$

Where $\Delta\widetilde{H}_{tot}$ expresses the total difference between idealized and empirical distribution (summation all over the *P* grid points)

The structural changes of the molecule during the LS step of the protein folding process can be simplified to the form shown in Fig.1.

The relation between the "*fuzzy-oil-drop*" and "oil-drop" model is presented in Fig 2 showing the difference between discrete model and the continuous fuzzy model.

The ES structural conformation is characterized by the large size and low packing. The size of "*fuzzy-oil-drop*" as defined for ES intermediate is much too large (packing density is much to low). This is why the size of "*fuzzy-oil-drop*" is decreased step-wise during the optimization procedure causing the squeezing of the molecule (Fig.3.) with simultaneous increase of hydrophobicity density in the central part of molecule.

The large $\sigma$ with low value of function's maximum, decrease of $\sigma$ causes the increased hydrophobicity density in a central part of the "drop" and high maximal density due to the small size expressed by $\sigma$

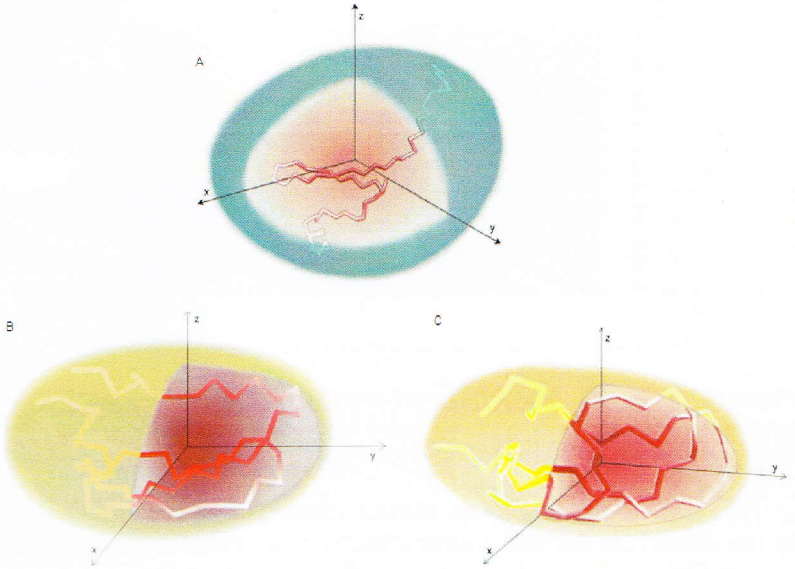


**Figure 1.** Schematic presentation of the "*fuzzy-oil-drop*" size decrease with the simultaneous increase of hydrophobicity density (red color intensity represents the change of hydrophobicity density increase).

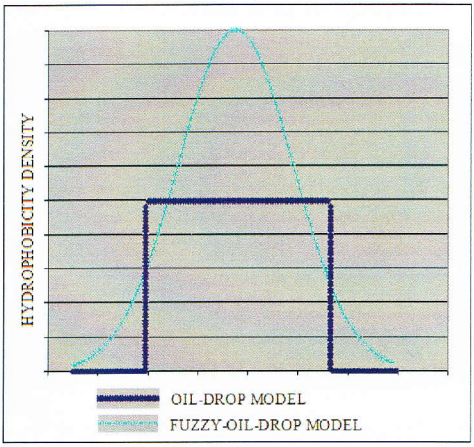


**Figure 2.** Traditional "oil-drop" model in respect to "*fuzzy-oil-drop*" model.

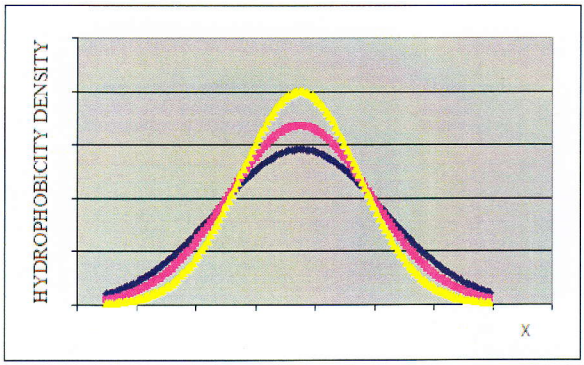


**Figure 3.** The visualization of "*fuzzy-oil-drop*" size change during the optimization procedure shown for one dimensional Gauss function influencing the hydrophobicity density increase in the central part of the molecule.

The surface under the curve remains constant (as the constant and equal to the value 1. is the total hydrophobicity) what makes proper the mutual relation between size and hydrophobicity density distribution.

What is the limit for the drop size decrease during the optimization procedure ?

The answer to this question was given using the analysis of the proteins molecules present in PDB in relation to their size in form of ES. All proteins of 150 aa or less in polypeptide (one chain molecules) were characterized by $\sigma_{xN}$, $\sigma_{yN}$, $\sigma_{zN}$ (N denotes the native structure) according to the procedure presented above. The same calculation was performed to the proteins in their ES structural forms (as described in [25]). The Phi, Psi angles of the proteins under consideration were "moved" on Ramachandran map to the nearest point belonging to ellipse path (limited conformational sub-space). The $Phi_e$, $Psi_e$ dihedral angles allowed generation of ES intermediate structure. The $\sigma_{xE}$, $\sigma_{yE}$, $\sigma_{zE}$ (index E denotes the ES intermediate) were calculated according to the same procedure as the $\sigma_{xN}$, $\sigma_{yN}$, $\sigma_{zN}$. The relation between the sizes for Native and ES structural forms is shown in Fig.3.

The approximation function (Fig.4.) expressing the dependence of V (volume of drop) on length (N – number of amino acids in polypeptide chain) (in logarithmic scale) has been found as follows:

$$\log V = 3.5671 + 0.7725 \times \log N \quad \text{for structures in native form}$$

$$\log V = 3.0013 + 1.2271 \times \log N \quad \text{for structures of ES intermediate}$$

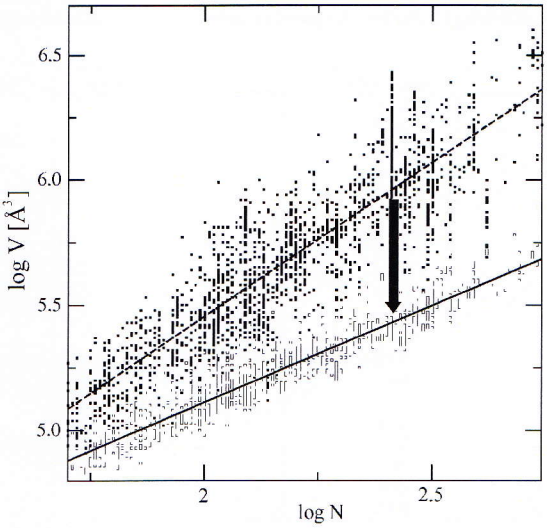


**Figure 4.** Relation between length of polypeptide (N) and volume of the drop (V) in logarytmic scale. The black squares – the ES intermediate, the gray squares – the N native structural form. The approximation lines: dashed – ES intermediate, solid – the native form expressing the relation between the polypeptide chain length and the size of molecule.

The correlation coefficients for ES and N structural forms are as follows: 0.95 and 0.88 respectively. The relation of drop size expressed by Dz : Dx : Dy (D – one dimensional measurement) has been found as follows: 1.00 : 0.85±0.08:0.79±0.09 for structures of native form and 1.00:0.67±0.14:0.53±0.12 for ES structural forms.

The two approximation functions represent the limits to which the optimization procedure shall be continued (Fig.4).

As was mentioned above, any hydrophobic scale available in the literature can be applied to express the empirical distribution of hydrophobicity. However the scale based on the "*fuzzy oil drop*" model can be also calculated. The one-domain proteins of the length 150 aa and below 150 aa were extracted from PDB. All proteins satisfying these conditions were taken to analysis. The molecules were oriented in coordinate system as described above. The localization of amino acids in particular areas (zones) of "*fuzzy-oil-drop*" determined the hydrophobicity of each amino acid (Tab.1.).

**Table 1.** Hydrophobicity scale for amino acids as estimated basing on the "*fuzzy oil drop*" model according to the localization of particular amino acid in particular place in the drop body.

| Amino acid | Hydrophobicity |
|---|---|
| LYS | 0.000 |
| ASP | 0.108 |
| GLU | 0.126 |
| GLN | 0.215 |
| PRO | 0.233 |
| ASN | 0.256 |
| ARG | 0.265 |
| SER | 0.314 |
| THR | 0.422 |
| GLY | 0.435 |
| ALA | 0.552 |
| HIS | 0.655 |
| TYR | 0.655 |
| MET | 0.825 |
| LEU | 0.834 |
| TRP | 0.874 |
| VAL | 0.892 |
| ILE | 0.942 |
| PHE | 0.982 |
| CYS | 1.000 |

The hydrophobicity scale generated according to this procedure appeared to be highly compatible with other hydrophobicity scales determined according to theoretical as well as empirical observations [4, 36-38] comparison of newly introduced scale with other ones is shown in Fig.5.

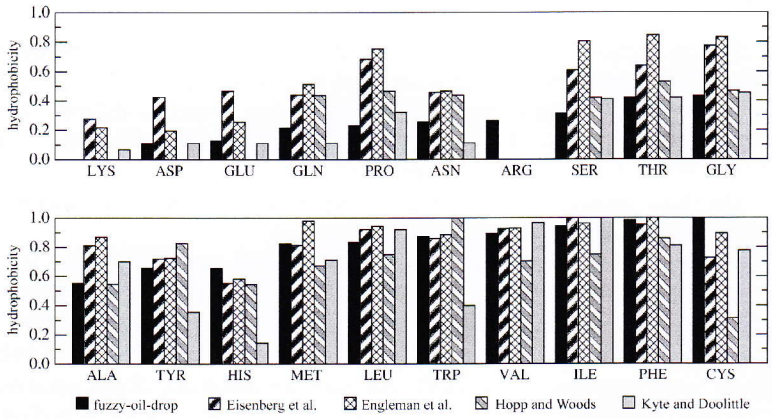


**Figure 5.** Comparison of different hydrophobicity scales and the one based on "*fuzzy oil drop*". The values are given in the standardized form (the range 0-1).

The procedure described so far concerns only the hydrophobicity driven part of the optimization procedure. Each step of such optimization is followed by the standard non-bonding interaction optimization.

In summary the procedure takes the following form:

I - starting point – the ES structural form of protein
II - traditional energy optimization (vdW interaction, electrostatic interaction, torsional potential)
III - hydrophobicity driven part of optimization procedure

1. orientation of the molecule in the coordinate system
2. simplification of site chains representation to effective atoms (geometric center of the side chain)
3. calculation of the size of "*fuzzy-oil-drop*"
4. generation of grid system
5. calculation of $\Delta\tilde{H}_i$
6. return to all atom representation (the atoms present in side chains re calculated for the effective atom position)
7. free rotations performance

IV- convergence condition (and final size of the molecule)

If hydrophobicity accordance satisfied – return to the point II for decreased size of drop.

If hydrophobicity accordance not satisfied – the free rotations performed and return to point III.

If hydrophobicity accordance satisfied and final size of drop achieved – long traditional energy minimization performed.

The procedure called III is performed in iteration procedure for gradually decreased size of "*fuzzy-oil-drop*" until the final size is obtained (see Fig.4.).

## The proteins folded according to "*fuzzy-oil-drop*" model

The small molecule (53 aa) used in CASP6 as the target TA0354_69_121 was taken as the first molecule to be folded according to the "*fuzzy-oil-drop*" model. The final structure is shown in Fig.6. The RMS-D versus the crystal structure is equal 5.656 Å. The non-bonding contacts shown in Fig. 7 reveal quite high similarity for native and LS structural form.

The comparison of structures (Fig.6.) and the distribution of non-bonding contacts (Fig.7.) seem to suggest that the "*fuzzy-oil-drop*" model produces

the promising results. Additional the final interpretation of the results will be also discussed in the Chapter VI.

Other examples of proteins folded according to the model under consideration are the following: ribonuclease, BPTI, lysozyme and both chains of human hemoglobin.

The characteristics of the size of example proteins is given in Tab. 2.

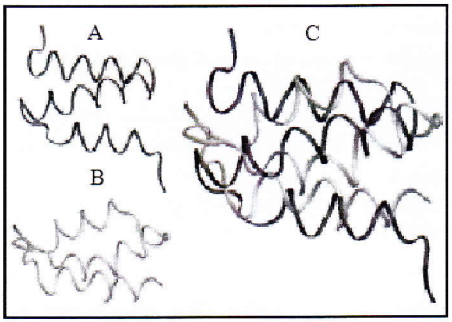


**Figure 6.** Ribbon presentation of the TA0354_69_121 protein – the LS structure overlapped together with native form. A – native form, B – folded *in silico* according to "*fuzzy-oil-drop*" model, C – two structures overlapped.

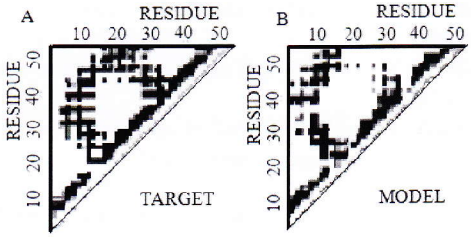


**Figure 7.** The non-bonding contacts for target and model structural forms of TA0354_69_121.

**Table 2.** Geometric characteristics of proteins taken as examples.

| Protein | ES size [Å] | | | N size [Å] | | | Size estimated [Å] | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | Y | Z | X | Y | Z | X | Y | Z |
| BPTI | 68.7 | 47.6 | 36.5 | 55.7 | 41.1 | 37.7 | 50.2 | 42.7 | 39.7 |
| Ribonuclease A | 143.4 | 63.6 | 59.6 | 61.8 | 50.2 | 48.4 | 61.1 | 51.9 | 48.2 |
| Lisozyme | 80.3 | 70.5 | 45.6 | 63.6 | 49.5 | 44.1 | 61.7 | 52.4 | 48.7 |
| Hemoglobin $\alpha$ | 91.0 | 63.0 | 56.7 | 59.4 | 56.7 | 45.4 | 63.1 | 53.6 | 49.9 |
| Hemoglobin $\beta$ | 108.3 | 57.1 | 56.7 | 63.6 | 60.6 | 42.1 | 63.7 | 54.1 | 50.3 |

# BPTI

The structures of BPTI (PDB ID – 4PTI [39]): ES structural form, N-native structure and the LS structural form as received applying the procedure mimicking the folding in the presence of external force field of hydrophobic character are shown in Fig. 8. No SS-bonds were defined during the simulation (no particular constraints ensuring SS-bonds generation were present in the procedure).

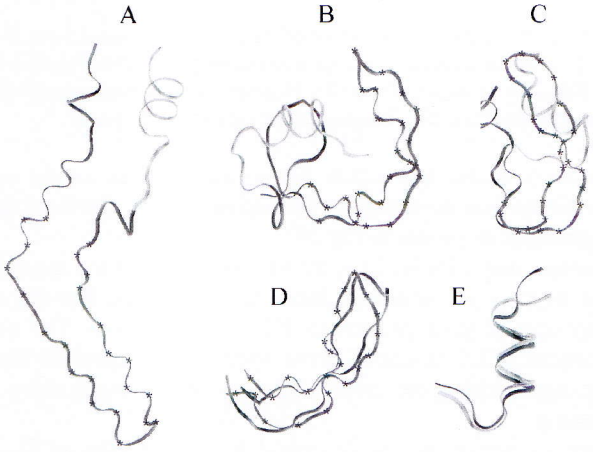


**Figure 8.** Models of BPTI: A – starting structure (ES), B – LS structural form, C – crystal structure. The secondary fragments of BPTI (distinguished as gray, black and with stars: central $\beta$-hairpin (13-38 - stars) and C-terminal $\alpha$-helix (44-58 - gray) D – the $\beta$-hairpin (native – gray, LS – black), E - helix fragments overlapped LS *in silico* (grey) and structure of native form (black).

Fig.8. visualizes the structural similarities of different structural forms of BPTI. The RMS-D calculated for ES and LS structural forms is equal 13.60 and 10.22 respectively although selected fragments represent much higher similarity.

The backbone Phi, Psi angles are localized well inside the allowed regions of the Ramachandran map (Fig.9.). The plot representing the distribution of Phi, Psi angles shows the proper migration of central $\beta$-sheet residues in the direction of C7eq during the LS folding simulation, while the C-terminal fragment keeps mostly right-handed helical form. The distribution of Phi, Psi angles as observed in LS form is the result of hydrophobicity driven optimization together with traditional internal energy minimization.

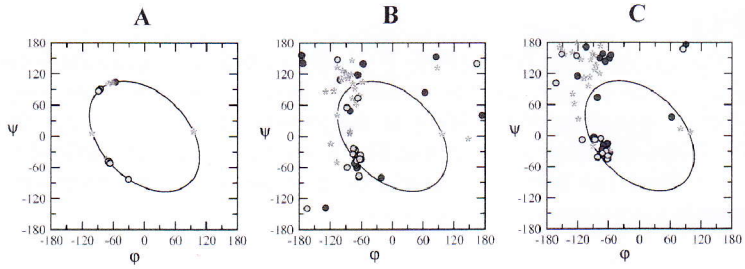


**Figure 9.** The φ, ψ angles distribution of BPTI: A – ES structural form, B – effect of LS simulation, C – crystal structure. Amino acids belonging to β-hairpin (13-38 - stars) and C-terminal fragment of α-helix form (44-58 gray) distinguished respectively. The ellipse path is shown to represent the limited conformational sub-space.

The residue-residue interaction can be traced based on the contact maps analysis. It seems that the approach toward the native distribution of contacts in LS is significant as shown in Fig.10.

The contact maps for BPTI seem to visualize well the similar to N form distribution even in ES structural form. It can be seen that the main strong stabilization core is even present in ES structural form. The new contacts which appeared in LS structural form appeared to represent the very good progress in approaching the correct contacts net although some interactions are still missing.

Another similarity can be described by the profiles of $D_{center\text{-}C\alpha}$ atoms (distance between geometric center and sequential Cα atoms) reveal some fragments of parallel mutual orientation what suggests proper mutual spatial orientation (Fig.11.).

The analysis of the profiles shown in Fig.11. allows distinguish the fragments of proper tendency to approach the orientation observed in crystal structure (fragments: 7-33 aa and 47-58 aa) high accordance of spatial orientation as well as those which are different in comparison to the native structural form.

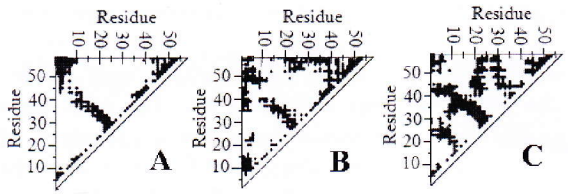


**Figure 10.** Contact maps as appeared in BPTI in ES form (A), LS form (B) and native crystal structure (C)

Other parameters: RMS-D, R-R (number of side chain-side chain contacts), ASA (solvent accessible area) and R radius of gyration calculated for the ES and LS structural forms of BPTI in respect to the native structure of proteins given in Tab.3. show also the approach to the native form.

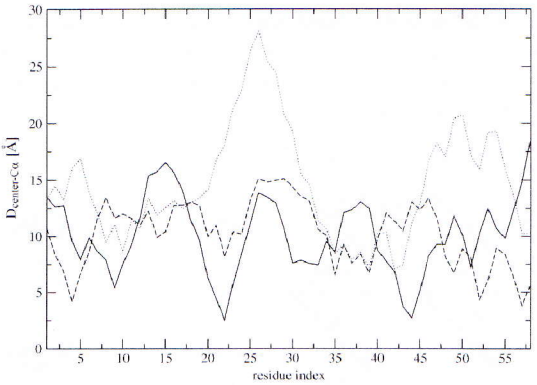


**Figure 11.** The profile of $D_{center-C\alpha}$ of BPTI in different structural forms: ES (dotted line), LS (dashed line) and native structure (solid line).

**Table 3.** The characteristics of the BPTI structural forms under consideration. The values called as R-R represent the number of side chains contacts reproduced in ES and LS forms. The parameters called $ASA_{tot}$ (Å$^2$), $ASA_P$ (Å$^2$) and $ASA_H$ (Å$^2$) measure the solvent accessible area: total, polar part and hydrophobic respectively. The parameters $R_g$ (Å) and $R_h$ (Å) measure the radius of gyration and hydrodynamic radius respectively.

| **Conformational state** | **RMSD [Å]** | **R-R** | **$ASA_{tot}$ [Å$^2$]** | **$ASA_P$ [Å$^2$]** | **$ASA_H$ [Å$^2$]** | **$R_g$ [Å]** | **$R_h$ [Å]** |
|---|---|---|---|---|---|---|---|
| Early-stage (*in silico*) | 13.60 | 270 | 5066.55 | 1883.83 | 3182.72 | 16.31 | 18.06 |
| Late-stage (*in silico*) | 10.22 | 419 | 4460.14 | 1599.97 | 2860.17 | 12.12 | 15.83 |
| Native | | 470 | 4018.73 | 1932.37 | 2086.36 | 12.41 | 16.11 |

The detailed analysis of BPTI folding *in silico* in the presence of external hydrophobic force field is presented in [40].

**Ribonuclease A**

The hydrophobic collapse simulation (according to "*fuzzy-oil-drop*" model) has been applied to ribonuclease A folding (PDB ID – 5RAT [41]). The early-stage folding has been shown in [42]. The ES structural form was taken as starting structure for "*fuzzy-oil-drop*" model to simulate the generation of LS intermediate.

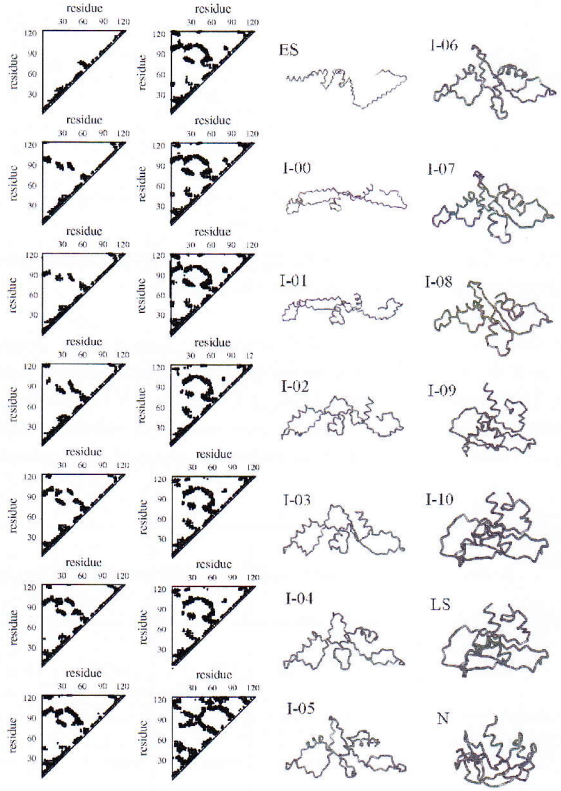


**Figure 12.** The contact maps and appropriate 3-D structures of ribonuclease A representing the sequential steps occurring during the LS simulation of ribonuclease A. ES – starting structure for LS simulation, I00-I10 – intermediates for sequential steps of decreasing size of drop, LS – post-simulation structure, N – native structure (according to PDB).

The monitoring of folding simulation of ribonuclease A is presented in Fig. 12. The ES form is characterized by the short-range contacts limited almost solely to backbone contacts. During the folding simulation the number of inter-residual contacts increased gradually leading to contacts suggesting the higher packing. The steps are presented and distinguished by I with sequential number in Fig.12.

The final contact map as well as the 3-D structural form appeared not to be satisfactory in comparison with the native structure.

The particular steps (I) seem to represent the process of hydrophobic collapse producing the higher packing density for each step (I).

The similarity of structures when the distance between the geometrical center and sequential Cα atoms ($D_{center\text{-}C\alpha}$) for structural forms under consideration reveals quite significant approach to the native structural form for simulation with the ligand present (Fig.13.).

The profiles of $D_{central\text{-}C\alpha}$ distance in native, ES and LS structural form of ribonuclease A (shown on Fig.13.) reveal quite good agreement particularly in a central and C-terminal fragment of the polypeptide chain (50-126 aa). The progress in approaching the profile characteristic for native structure producing the quite similar for LS structural form makes the model promising. The reason for the presence of discrepancy will be explained in the Chapter 6 of this book.

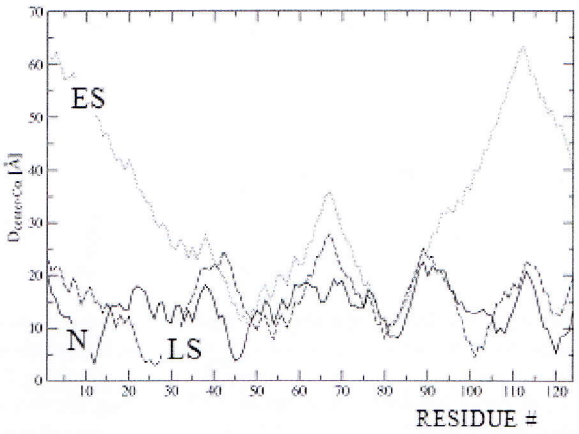


**Figure 13.** Profiles of $D_{center\text{-}C\alpha}$ for ribonuclease A : solid line – the crystal form, dashed line – the for simulation with the ligand present and dotted line for ES intermediate (the starting point of simulation).

Other characteristics of the process of hydrophobic collapse is given in Fig. 14. The decrease of drop size seems to induce increase of the number of non-bonding contacts reaching the number close to the native one (Fig.14.A.). The change of radius of gyration is shown in Fig.14.B. The detailed discussion of the value of this parameter is shown in Tab. 4. The 40.84% of all non-bonding (1304) contacts in LS structure appeared to be accordant with the native one.

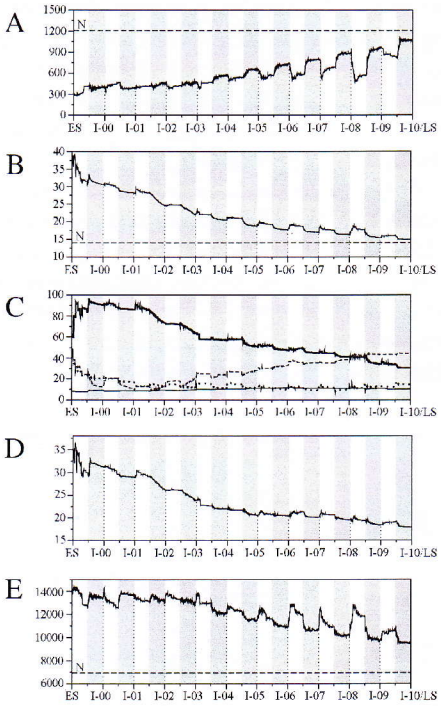


**Figure 14. The changes during simulation: ribonuclease A. A.** number of non-bonding contacts (NB). **B.** radius of gyration (Rg Å),solvent accessible surface (ASA (A) C – RMS-D calculated for Cα atoms versus the native structure, E – accessible surface area **C.** distance between Cα atoms of CYS creating the disulphide bonds: C26-C86 – solid line, C40-C95 – dashed line, C58-C110 – dotted line, C65-C72 – dashed/dotted line (values shown in [Å]). **D.** RMS-D – Cα vs native structural form [Å]. **E.** Accessible surface area [Å]. The horizontal lines denoted by N represent the level of particular variable as appeared in native form.The vertical white rectangles represent the step of traditional non-bonding optimization, the gray rectangles represent the hydrophobicity driven step of optimization procedure.

All distances related to SS bonds creation except C65-C72 were found to be in large mutual distance in the ES especially between C58 and C110. The significant mutual approach has been observed for C26-C84 and C58-C110 as can be seen in Fig.14.C.

The global RMS-D decreased during LS folding simulation from the value 32.27A for ES to 18.07 A for LS. This value is not satisfactory although the explanation for this large value will be presented in the next chapter. The gradual decrease of RMS-D value during hydrophobic collapse is shown in Fig.14. D.

The size of accessible solvent area decreases also gradually during simulation although the final structure represents higher than expected value of this parameter (Fig.14.E).

All the profiles shown in Fig.14. demonstrate the character of the process which may be approached to the squeezing step (decrease of drop size) followed by the relaxation (procedure expressing the traditional optimization minimizing the internal energy of the system). The non-shaded fragments on the Fig. 14. represent the hydrophobicity driven part of optimization procedure.

The comparison of Rg of ribonuclease A makes the resultant structure helps treat the model as promising.

**Table 4.** The values of radius of gyration (Rg) for ribonuclease A according to simulations and experiments. a – Sosnick et al. [43], b – Zhou et al. [44], c – calculated, d – simulation of LS, e – ES structural form, f – theoretical value given by Zhou [45] according $6R^2_g = 130n$ (n=number of amino acids in polypeptide) with the correction introduced by Tanford [45]

| STRUCTURE | $R_g$ [Å] |
|---|---|
| Native | 15.0[a]<br>15.4[b]<br>14.1[c] |
| LS | 15.0 |
| Thermal denaturation[a]<br>-S-S- bonds declared<br>-S-S- bonds reduced | <br>19.3<br>28.0 |
| Chemical denaturation[b]<br>-S-S- present<br>-S-S- reduced | <br>17.3<br>24.0 |
| ES[e] | 36.6 |
| Entirelly random[f] | 45 |

## Lisozyme

The next molecule taken as the example to verify the applicability of "*fuzzy-oil-drop*" to simulate the LS step of protein folding process is lisozyme (PDB ID – 2EQL [46]).

The structural forms: ES, LS and native one are shown in Fig.15 together with the non-bonding contact maps. They reveal that the number of contacts is lower than the expected one although the progress of increasing the interaction seems to be accordant to expectations. The detailed analysis of lisozyme structural forms in relation to "*fuzzy-oil-drop*" model is presented in [47].

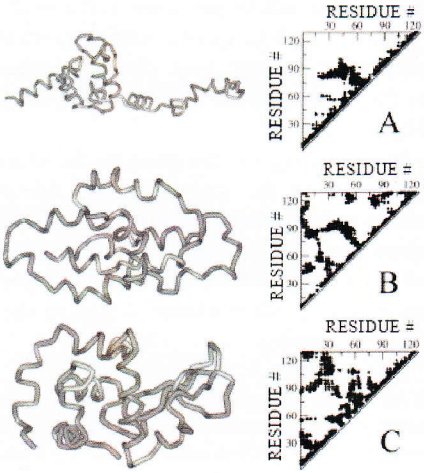


**Figure 15.** The structure of ES (A), LS (B) and native form (C) of lisozyme with the appropriate contact maps.

**Table 5.** Different types of hydrogen bonds involved in secondary structure formation for all discussed structural forms of lysozyme. The given values are calculated per 100 residues.

| **Conformational state** | **Hydrogen bond type** | | | | | |
|---|---|---|---|---|---|---|
| | parallel bridge | antiparallel bridge | O($i$)→H-N($i$+2) | O($i$)→H-N($i$+3) | O($i$)→H-N($i$+4) | O($i$)→H-N($i$+5) |
| Early-stage | 1.6 | 0.0 | 14.7 | 3.1 | 37.2 | 4.7 |
| Late-stage | 0.0 | 0.0 | 10.9 | 24.0 | 27.1 | 1.6 |
| Native | 1.5 | 9.2 | 7.7 | 20.0 | 26.2 | 2.3 |

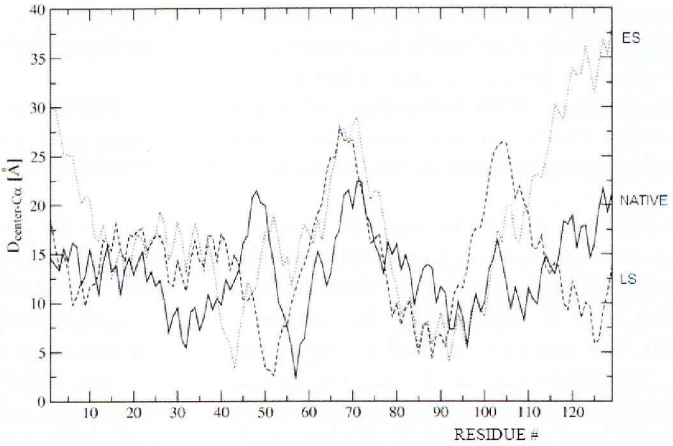


**Figure 16.** The profiles of $D_{center-C\alpha}$ distances in lisozyme – crystal form (solid line), ES (dotted line), LS (dashed line)

The summary presenting the H-bonds presence in lisozyme in different structural forms are given in Tab. 5. The H-bonds as present in $\beta$-structural forms: parallel and antiparallel do not appear in the model structure. The $\beta$-conformation generally is difficult to be predicted using any methods applied for protein structure prediction [48].

The profiles of $D_{center-C\alpha}$ distances in lisozyme shown in Fig. 16. reveal fragments of quite high accordance although some discrepancies can be seen (45-55 aa, C-terminal fragment starting at 115 aa).

# Conclusions

The "*fuzzy-oil-drop*" model was applied to simulate the environment of hydrophobic character for folding process. The molecule is directed to localize the hydrophobic residues in the central part of the molecular body with the hydrophilic residues exposed on the protein surface. The molecule folded according to such procedure appears to be entirely covered by the hydrophilic residues. The consequences of such hydrophobicity distribution make the protein molecule very well soluble. The disadvantage of the structure generated according to the presented procedure is that the molecule does not represent any possibility to interact with other molecules. In consequence one may say the resulted molecule is of no biological activity. The comparison of the LS structural form with the crystal structure reveals the lack of the specific ligand binding site.

The conclusion based on this observation is that it is necessary to analyze the hydrophobicity distribution in proteins in their crystal form. This analysis will be shown in the next chapter of this book.

The summary of the applicability of "*fuzzy-oil-drop*" model and its place in the entire folding process simulation *in silico* (including the ES step and the databases implemented in this procedure) can be presented as shown in Fig.18.

The left site shows the analysis oriented on the search for early-stage parameterization based on the crystal structures of proteins. It represents sort of "step back" procedure starting from the native structure (A) and its partial unfolding what is expressed as the change to limited conformational sub-space(B). The structure obtained in consequence of such procedure is shown in (C). The distribution of $Phi_e$ and $Psi_e$ (e stands from ellipse path) dihedral angles along the ellipse path appeared to be characteristic for each amino acid (D). Seven local maxima can be distinguished in the probability distribution profile (D). Each of them may be called according to the letter codes as shown in (D). The structure of ES intermediate of particular protein can be coded as shown in (E). The comparison of amino acid sequence and letter codes for structural form may be used to construct the sequence-to-structure contingency table expressing the dependence between these two characteristics. The contingency table has been generated for tetrapeptides as units (in overlapping system).

The structure of particular protein can be generated following the procedure shown in right part of the Fig. 17. The letter codes expressing structural forms for particular amino acid sequence can be attributed to define the structure of ES on the basis of contingency table generated according to "step-back" procedure (F).

In contrast to the probability distribution which for "step-back" procedure is of continuous form, the distribution in folding procedure is of discrete form. Thus the $Phi_e$, $Psi_e$ dihedral angles for local maximum position are taken to define the structure of ES intermediate (G), which is shown in (H). This structural form is treated as staring point for LS simulation. The "*fuzzy-oil-drop*" model based procedure (I) applied for LS folding produces the structure as shown in (J).

The "*fuzzy-oil-drop*" model applied to fold the LS step of the folding simulation procedure is described in this chapter. The general conclusion is that the LS folding produces the structures which are not of satisfactory high similarity to the native structure although some characteristics of final structure seems to be promising.

The modification of the "*fuzzy-oil-drop*" model to make it more reliable is shown in next chapter of this book.

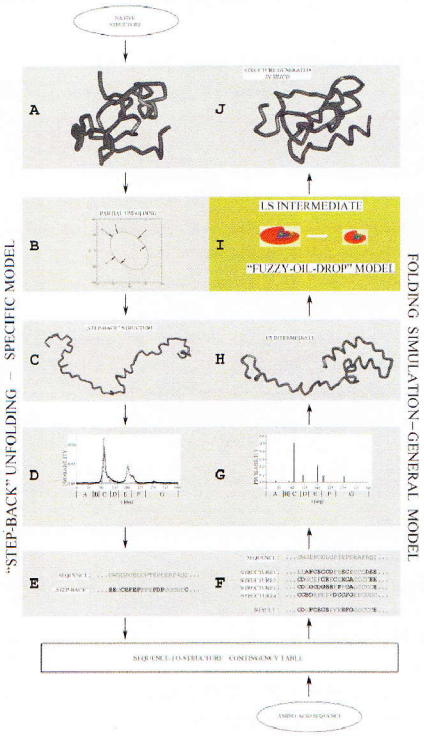


**Figure 17.** The summary of the folding process simulation, the part of which is the LS intermediate generation. A to E – "step-back" procedure expressing the partial unfolding. The native structure (A) partially unfolded to the ES intermediate changing the original Phi, Psi angles by moving them toward the ellipse (B) appears to be as presented in (C). The movement of Phi Psi angles toward ellipse path produces the (continuous) probability profile along the ellipse (D). The letter codes introduced to distinguish the local probability maxima allow structure description by identification of particular maximum. When all amino acids in all proteins present in PDB (nonredundant representation) are taken into consideration the characteristic profile for each amino acid can be generated (E). The folding simulation procedure (F to J) starts, when amino acid sequence is known. According to contingency table (sequence-to-structure expressed by letter codes) the sequence of structural letter codes can be attributed to the amino acid sequence under consideration (F). Unfortunately the letter code points out only the position of local maximum (discrete form) (G). Knowing the appropriate $Phi_e$, $Psi_e$ dihedral angles for each local probability maximum the structure as shown in (H) can be generated. The next step is to apply the LS step simulation based on "*fuzzy-oil-drop*" model (I). This procedure produces the structure as shown in (J).

# References


1. Kauzmann W. 1959 Adv Protein Chem. **14,** 1-63.
2. Klapper M.H. 1971 Biochim Biophys Acta. **229**(3), 557-66.
3. Klotz I.M. 1970 Arch Biochem Biophys. **138**(2), 704-6.
4. Kyte J, Doolittle R.F. 1982 J Mol Biol. **157**(1), 105-32.
5. Meirovitch H. Scheraga H.A. 1980 Macromolecules. **13**(6), 1398-1405.
6. Meirovitch H. Scheraga H.A. 1980 Macromolecules. **13**(6), 1406-1414.
7. Meirovitch H. Scheraga H.A. 1981 Macromolecules. **14**(2), 340-345.
8. Chothia C. 1975 Nature, 254, 304-308.
9. Rogov S.I, Nekrasov A.N. 2001 Prot. Eng. 14, 459-463.
10. Rose G.D, Roy S. 1985 Science 229, 834-838.
11. Schwartz R, Istrail S, King J. 2001 Protein Sci. 10, 1023-1031.
12. Braig, K., Otwinowski Z, Hegde R, Boisvert D.C, Joachimiak A, Horwich A.L, Sigler P.B. 1994 Nature, **371**(6498), 578-86.
13. Horwich, A.L., E.U. Weber-Ban, and D. Finley. 1999, Proc Natl Acad Sci U S A. **96**(20), 11033-40.
14. Horwich A. L, Weber-Ban E.U, Findey D. 1999, Proc. Natl. Acad. Sci. USA, 96, 11033-11040.
15. Houry, W.A., Frishman D, Eckerskorn C, Lottspeich F, Hartl F.U. 1999 Nature, **402**(6758), 147-54.
16. Sakikawa C, Taguchi H, Makino Y, Yshida M. 1999 J. Biol. Chem. 274, 21251-21256.
17. Wang, Z., Feng H, Landry S.J, Maxwell J, Gierasch L.M. 1999, Biochemistry, **38**(39), 12537-46.
18. Xu Z, Horwich A.L, Sigler P.B. 1997 Nature, 388, 741-750.
19. Baldwin, R.L., 2002 Science, **295**(5560), 1657-8.
20. Dill, K.A. 1990 Biochemistry, **29**(31), 7133-55.
21. Finney, J.L., Bowron D.T, Daniel R.M, Timmins P.A, Roberts M.A. 2003 Biophys Chem, **105**(2-3), 391-409.
22. Pace, C.N., Shirley B.A, McNutt M, Gajiwala K. 1996 FASEB J, **10**(1), 75-83.
23. Kurochkina, N. and G. Privalov, 1998 Protein Sci. **7**(4), 897-905.
24. Bonneau, R., C.E. Strauss, and D. Baker, 2001 Proteins, **43**(1), 1-11.
25. Roterman I. Brylinski M, Jurkowski W. In recent Advances In Structural Bioinformatics, 2007 Editor: Alexandre de Brevern, Research Signpost Kerala, India pp 69-104.
26. Roterman, I. 1995, J. Theoretical Biol. 177, 283-288.
27. Roterman, I., and Konieczny, L. 1995, Computers and Chemistry 19, 247.
28. Roterman, I. 1995, Biochimie, 77, 204-216.
29. Jurkowski, W., Brylinski, M., Konieczny, L., and Roterman, I. 2004, J Biomol Struct Dyn 22, 149-158.
30. Brylinski, M., Jurkowski, W., Konieczny, L., and Roterman, I. 2004. Bioinformatics 20, 199-205.
31. Brylinski, M., Jurkowski, W., Konieczny, L., and Roterman, I. 2004, TASK Quarterly 8, 413.

32. Brylinski, M., Konieczny, L., Czerwonko, P., Jurkowski, W., Roterman, I. 2005, J. Biomed. Biotech. 2, 65-79.
33. Meus, J., Brylinski, M., Piwowar, M., Piwowar, P., Stefaniak, J., Wisniowski, Z., Jurkowski, W., and Roterman, I. 2005 Med. Sci. Monit. 12(6), BR208-214.
34. Konieczny L, Brylinski M, Roterman I. 2006 In Silico Biol. 6, 15-22.
35. Levitt M.1976. J Mol Biol, **104**(1): 59-107.
36. Hoop T.P, Woods K.R. 1981. Proc. Natl. Acad. Sci. USA 78, 3824-3828.
37. Eisenberg D, Schwarz E, Komaromy M, Wall R 1984. J. Mol. Biol. 179, 125-142.
38. Engelman D.M, Steitz T.A, Goldman A. 1986. Annu Rev Biochem Biophys Chem 15, 321-353.
39. Marquart M, Walter J, Deisenhofer J, Bode W, Huber M. 1983. Acta Cryst B, 39, 480-490.
40. Brylinski M, Brylinski M, Konieczny L, Roterman I. 2006 Biochimie 88 1229-1239.
41. Tilton R.F, Jr, Dewan J.C, Petsko G.A. 1992 Biochemistry 31, 2469-2481.
42. Brylinski M, Konieczny L, Roterman I. 2006 Computational Biol and Chem. 30: 255-267
43. Sosnick T.R, Trewhella J. 1992 Biochemistry 31, 8329-8335.
44. Zhou J.M, Fan Y.X, Kihara H, Kimura K, Amemiya Y. 1998. FEBS Lett 430, 275-277.
45. Tanford C 1968. Adv. Protein Chem 23, 121-282.
46. Tsuge H, Ago H, Noma M, Nitta K, Sugai S, Miyano M. 1992. J. Biochem. 111, 141-143.
47. Brylinski M, Konieczny L, Roterman I. 2006 J. Biomol. Struct Dynam. 23: 519 - 527
48. Orengo C.A, Bray J.E, LoConte L, Sillitoe I. 1999, Proteins Structure, Function and Genetics 37, 149-170.