Is the protein folding an aim-oriented process? Human haemoglobin as example

Michal Brylinski

Department of Bioinformatics and Telemedicine,
Collegium Medicum, Jagiellonian University,
Kopernika 17, 31-501 Krakow, Poland

Faculty of Chemistry, Jagiellonian University,
Ingardena 3, 30-060 Krakow, Poland
E-mail: mybrylin@cyf-kr.edu.pl

Leszek Konieczny

Institute of Medical Biochemistry,
Collegium Medicum, Jagiellonian University,
Kopernika 7, 31 034 Krakow, Poland
E-mail: mbkoniec@cyf-kr.edu.pl

Irena Roterman*

Department of Bioinformatics and Telemedicine,
Collegium Medicum, Jagiellonian University,
Kopernika 17, 31-501 Krakow, Poland

Faculty of Physics,
Astronomy and Applied Computer Science,
Jagiellonian University, Ingardena 3, 30-060 Krakow, Poland
E-mail: myroterm@cyf-kr.edu.pl
*Corresponding author

Abstract: The model for protein folding (*in silico*) simulation is presented. Three steps have been implemented:

- early stage folding based on the backbone conformation
- hydrophobic collapse based on the *fuzzy-oil-drop* model
- aim-oriented structure modification by the function-related ligand.

The model has been verified taking α and β haemoglobin chains as examples to fold them in two different conditions: with and without haem being present in the folding environment. The presence of haem and its participation in the folding simulation led to the structure more similar to the crystal one. It suggests that the haem presence directs the folding process towards the function-related structure.

Keywords: hydrophobic collapse; protein folding; protein structure prediction; active site; protein–protein complex; ligand binding; haemoglobin; hydrophobic scale; oil-drop model; ligand library; bioinformatics.

Reference to this paper should be made as follows: Brylinski, M., Konieczny, L. and Roterman, I. (2007) 'Is the protein folding an aim-oriented process? Human haemoglobin as example', *Int. J. Bioinformatics Research and Applications*, Vol. 3, No. 2, pp.234–260.

Biographical notes: Michal Brylinski graduated from Medical University in Wroclaw School of Pharmacy with an MSc Degree in Pharmacy and from the Faculty of Chemistry, Jagiellonian University in Cracow, with a PhD Honours Degree in Chemistry. He is a Research Associate in the Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College. His research interests are generally focused on protein structure, folding and dynamics. Specific areas of current research are ab initio folding, protein–ligand interactions and large-scale simulations of biomolecules.

Leszek Konieczny is a Professor Emeritus in the Institute of Medical Biochemistry, Collegium Medicum, Jagiellonian University Krakow, Poland. He was educated in medicine with a MD in 1964 and Habilitation in 1979. He is a specialist in experimental medical biochemistry and molecular immunology. His area of interest is systems biology.

Irena Roterman is a Professor of Bioinformatics, Head of the Department of Bioinformatics and Telemedicine Collegium Medicum, Jagiellonian University, Krakow, Poland. She collaborates with the Faculty of Physics, Jagiellonian University, Krakow, Poland. She did her MSc in Chemistry 1974, PhD in Natural Sciences in 1985, Habilitation in Biochemistry in 1995. She has been a Professor in Biochemisry since 2005. She is interested in protein structure prediction and computer-aided drug design.

1 Introduction

The classic work by Kauzmann (1959) emphasised hydrophobic interactions as crucial for formation and stabilisation of the protein tertiary structure (Klapper, 1971; Klotz, 1970; Meirovitch and Scheraga, 1980a, 1980b). Hydrophobic interactions are treated as responsible for hydrophobic core creation, with hydrophilic residues localised on the protein surface (Kyte and Doolittle, 1982; Meirovitch and Scheraga, 1981). The amino-acid sequence partitions a protein into its densely packed interior with the local hydrophobicity maximum and surface area of less densely packed sites of local minimum of hydrophobicity (Rose and Roy, 1980). Spatial distribution of amino-acid hydrophobicity was even used to differentiate native and non-native protein structures (Baumann et al., 1989; Bonneau et al., 2001; Holm and Sander, 1992; Novotny et al., 1988). Detailed analysis of the spatial variation of hydrophobicity, focused on the region of transition between protein interior and exterior, was carried out for 30 relatively diverse globular proteins as well as for 14 decoys (Silverman, 2001). Apart from soluble proteins, the distribution of apolar and polar residues provided comprehensive information about transmembrane protein architecture (Eisenberg et al., 1984; Engelman and Zaccai, 1980; Rees et al., 1989; Silverman, 2003). Furthermore, the hydrophobic effect as a dominant driving force in protein folding was suggested (Baldwin, 2002; Dill, 1990; Finney et al., 2003).

The *Critical Assessment of Predicted Interaction* (CAPRI) is oriented on blind prediction of protein complexes with ligands and protein–protein complex creation (Janin, 1979, 2005a, 2005b, Janin et al., 2003; Mendez et al., 2003, 2005). The results of competitions are presented on a website (http://capri.ebi.ac.uk) and additionally published in Proteins Struc Func Gen (2003) 52 (the whole volume is devoted to this problem). The papers presented there present methods aimed on complex creation, based mostly on the detailed geometric surface analysis aimed at the irregularity and cavity search.

Some of the techniques used to identify functionally important residues from sequence or structure are based on searching for homologues proteins of known function (Bork et al., 1998; Skolnick and Fetrow, 2000; Wallace et al., 1997; Zvelebil and Sternberg, 1988). However, homologues, particularly when the sequence identity is below 25%, need not have related activities (Devos and Valencia, 2000; Hegyi and Gerstein, 1999; Wilson et al., 2000). Geometry-based methods have shown that the location of active-site residues can be identified by searching for cavities in the protein structure (Liang et al., 1998) or by docking small molecules onto the structure (Oshiro et al., 1995). The cave localisation *in silico* has been presented on the basis of the characteristics of the normal created for each surface piece (Lamb et al., 2001). The complex analysis of protein interfaces and their characteristics vs. the highly divergent areas is presented in Jimenez (2005). Several experimental studies have shown that mutation of residues involved in forming interfaces with other proteins or ligands can also be replaced to produce more stable, but inactive, proteins (Kanaya et al., 1996; Meiering et al., 1992; Shoichet et al., 1995; Zhang et al., 1992). On this basis, several effective algorithms were developed (Elcock, 2001; Ondrechen et al., 2001). Structural analysis, coupled with measures of surface hydrophobicity, has been used to identify sites on the surfaces of proteins involved in protein–protein interactions (Jones and Thornton, 1997).

The model presented in this paper oriented on localisation of area responsible for ligand binding or protein–protein complex creation is based on characteristics of spatial distribution of hydrophobicity in a protein molecule. It is assumed that hydrophobicity changes from protein interior (maximal hydrophobicity) to exterior (close to zero level of hydrophobicity), according to the 3-dimensional Gaussian distribution. However, it is generally accepted that the core region is not well described by a spheroid of buried residues surrounded by surfaces residues due to hydrophobic channels that permeate the molecule (Kuntz and Crippen, 1979; Crippen and Kunz, 1978). Therefore, the simple comparison of theoretical (idealised according to Gaussian function) and empirical spatial distribution of hydrophobicity in protein gives the opportunity to identify the regions with high deviation vs. the ideal model. Those regions recognised by high hydrophobicity density differences seem to reveal functionally important sites in proteins.

The consequence of above observation is that the 3-dimensional Gaussian function can play the external force field, directing the folding molecule to concentrate the hydrophobic residues in central part of molecule, with hydrophilic ones exposed to the surface (Brylinski et al., 2006a, 2006b, 2006c, 2006d; Konieczny et al., 2006). It can be reached during the optimisation procedure, in which the criterion for optimisation is the smallest possible difference between idealised and observed hydrophobicity distribution in an ellipsoid. Another consequence of the above observation is the active presence of natural ligand in the environment of folding molecule. The presence of external force field influences the polypeptide-folding process as well as the appropriate orientation of ligand. Both protein and ligand are directed by the external force field.

The model is presented using α and β haemoglobin chains as examples. Moreover, the paper presents the model for active site localisation and gives the proposal for protein structure prediction in the form of protein-folding simulation, which includes active presence of natural ligand in this process.

2 Materials and methods

2.1 Data

The proteins selected for analysis are goose lysozyme (154L) (Weaver et al., 1995), sperm whale myoglobin (1A6M) (Vojtechovsky et al., 1999), bovine acyl-coenzyme A (1ACA) (Kragelund et al., 1993), catalytic domain of human TNF-α converting enzyme (1BKC) (Maskos et al., 1998), human lysozyme (1LZR) (Song et al., 1994), bovine ribonuclease A (1RND) (Aguilar et al., 1992), dihydrofolate reductase from *E. coli* (3DRC) (Warren et al., 1991), human mitogen-activated protein kinase (1A9U) (Wang et al., 1998) and human cyclin-dependent kinase CDK6 complex with cell-cycle inhibitor P19INK4D (1BLX) (Brotherton et al., 1998). Some of these proteins have been analysed in respect of their biological function localisation, using molecular dynamics simulation (Izrailev and Farnum, 2004; Luque and Freire, 2000; Fermi et al., 1984) and electrostatic energy calculation (Elcock, 2001). This selection was aimed at making possible the comparative analysis of results, obtained by different groups and different methods. The counted proteins were used to verify the *fuzzy-oil-drop* model as a tool to recognise the active-site localisation in protein molecule.

The amino-acid sequences of α and β chains of human haemoglobin (PDB ID: 3HHB; Levitt, 1976) have been taken as the model to demonstrate the influence of *fuzzy-oil-drop* on the protein-folding process simulation. Two versions of simulations (for each chain) were applied: one for isolated polypeptide chain folding and the second in the presence of natural ligand – haem.

2.2 Empirical (observed) oil drop

The standard unified spatial orientation of protein molecule is necessary to calculate the empirical hydrophobicity distribution. The geometrical centre of the molecule is localised in the origin of coordinate system. The longest distance between effective atoms (side chains represented by geometrical centre of atoms present in side chain of amino acid) in protein molecule determines the *Z*-axis. The *Y*-axis is oriented according to the longest distance between projections of effective atoms on the *XY* plane.

The grid points localised in a positions of effective atoms of side chains are used to calculate the empirical hydrophobicity density, according to the simple sigmoid function previously proposed to quantitatively describe the hydrophobic interactions (Levitt, 1976). The *j*th point (which is the *j*th residue) collects hydrophobicity $\tilde{H}o_j$ as follows:

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

where $\tilde{Ho}_j$ represents the empirical hydrophobicity value characteristic for *j*th grid point (*j*th amino acid), H_i^r represents the hydrophobicity of the *i*th amino acid, r_{ij} – distance between the *j*th grid point (*j*th residue) and effective atom of *i*th amino acid and *c* – expresses the cut-off distance that has a fixed value of 9.0Å, following the original paper (Levitt, 1976). $\tilde{Ho}_{\text{sum}}$ represents the total hydrophobicity (sum of all grid points' hydrophobicity).

The application of this function requires the parameters of hydrophobicity attributed to each amino acid. Many scales for residue hydrophobicity are available. Some of them are based on analysis of known protein 3D structures (Kyte and Doolittle, 1982; Eisenberg et al., 1982; Engelman et al., 1986; Hopp and Woods, 1981; Rose et al., 1985), while the others are derived from the physicochemical properties of amino acid side chains (Wimley and White, 1996; Wolfender et al., 1981). The selection of appropriate scale seems to be a crucial issue; therefore, a new statistical-based hydrophobicity scale for amino acids has been created (Brylinski et al., 2006b) (description of the scale will be given later in this paper).

2.3 Theoretical (expected) hydrophobic core

The same grid points as described above are used to calculate the theoretical (expected) hydrophobicity distribution representing the idealised *fuzzy-oil-drop*, which can be applied to simulate hydrophobic collapse in protein-folding process. Hydrophobic core applied in this model is assumed to be represented by 3-dimensional Gaussian function:

$$\tilde{Ht}_j = \frac{1}{\tilde{Ht}_{\text{sum}}} \exp\left(\frac{-(x_j-\overline{x})^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_j-\overline{y})^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_j-\overline{z})^2}{2\sigma_z^2}\right). \tag{2}$$

$\tilde{Ht}_j$ is assumed to represent the hydrophobicity density in *fuzzy-oil-drop*. The hydrophobicity maximum localised in a centre of ellipsoid decreases in form of distance dependence according to 3-dimensional Gaussian function. The parameter – mean value, for which Gaussian function reaches its maximum (maximum of hydrophobicity density) is localised in (0, 0, 0) origin of coordinate system. The second parameter – standard deviation – represents the size of drop (the values of three standard deviations determines the size of drop: σ_x, σ_y, σ_z) and depends on the length of polypeptide under consideration (Brylinski et al., 2006a). The values of $\tilde{Ht}_j$ calculated for each grid point are standardised to give the value of sum of all values over all grid points equal to 1.0.

The analysis of ellipsoid size, as dependent on polypeptide chain length, has been presented in Brylinski et al. (2006a), where single-domain proteins present in PDB have been taken into account.

2.4 Ligand-binding site

Since both theoretical $\tilde{Ht}$ and observed $\tilde{Ho}$ distributions of hydrophobicity are standardised to 1.0 and the partition values are calculated for the same grid points, the comparison of these two distributions is possible. The differences between theoretical and empirical distributions $\Delta\tilde{H}$ express the irregularity of hydrophobic core construction. For *i*th residue $\Delta\tilde{H}_i$ is calculated as follows:

$$\Delta\tilde{H}_i = \tilde{H}t_i - \tilde{H}o_i \tag{3}$$

where $\tilde{H}t_i$ and $\tilde{H}o_i$ are the theoretical and observed values of hydrophobicity for the geometric centre of *i*th residue, respectively (what is called also as a grid point). The theoretical and empirical *fuzzy-oil-drop* was calculated for all proteins taken into consideration. The colour scale was introduced to express the magnitude of difference $\Delta\tilde{H}_i$ in particular protein area what enables visualisation of the localisation of these discrepancies in the protein molecule.

2.5 Folding simulation

The folding simulation is based on the minimisation of the difference between idealised and observed hydrophobicity density distribution. The procedure in a step-wise form is as follows:

- The size of the ellipsoid (expressed by σ_x, σ_y, σ_z) covering the early-stage folding structure is calculated for the orientation described above.
- The grid system is constructed for the spatial orientation of molecule according to description in *Empirical oil-drop*. The grid (covering the whole molecule) step has the fixed value of 5.0 Å in all simulations.
- The empirical hydrophobicity values are also calculated for each grid point (equation (1)).
- Theoretical, idealised hydrophobicity values are calculated for each grid point according to 3-dimensional Gaussian function (equation (2)).
- Optimisation procedure:
  - The structural changes are performed to approach the idealised hydrophobicity density making the virtual oil-drop represented by folding molecule more similar to the theoretical one. This procedure optimises $\Delta\tilde{H}_{\text{tot}}$:

    $$\Delta\tilde{H}_{\text{tot}} = \sum_{j=1}^{P} (\tilde{H}t_j - \tilde{H}o_j)^2 \tag{4}$$

    where Ht_j and Ho_j are the theoretical and observed values of hydrophobicity for the *j*th grid point, respectively. *P* denotes the total number of grid points.
  - The algorithm given by Rosenbrock (1960) was applied to optimise $\Delta\tilde{H}_{\text{tot}}$ during the simulation.
  - Each calculation of hydrophobicity distribution approach is followed by traditional energy minimisation procedure to avoid some forbidden overlaps. All-atom representation is adopted for these steps. The energy minimisation sub-step is performed according to ECEPP/3 standards (Sippl et al., 1984; Momany et al., 1975; Dunfield et al., 1978; Nemethy et al., 1983, 1992).

- The optimisation procedure is carried out as long as the assumed convergence condition is reached.
- The next step of folding simulation in iteration procedure is performed for the theoretical oil drop of smaller size than the previous one, to force the molecule to get more compact. The decrease of drop size causes in consequence the increase of hydrophobicity density in a central part of drop. The size relation between early-stage structure and the native one as dependent on number of amino acid in polypeptide chain has been published (Brylinski et al., 2006a).
- The procedure stops, when the final size of virtual oil drop reaches the size similar to the expected one and when the convergence in hydrophobicity and potential energy optimisation procedures is reached.
- For every 20 steps of applied procedure for structure optimisation, the coordinates were saved to monitor the structural changes during the folding process simulation.

The procedure can be interpreted as a two-step process (repeated iteratively): squeezing (hydrophobicity optimisation) followed by the relaxation of the structure (energy optimisation). The product of this procedure is a molecule covered mostly by the hydrophilic residues, with all hydrophobic ones buried in the interior of the molecule. In consequence, the received molecule seems to be very well soluble on one hand but very inactive on the other.

2.6 Presence of haem

The simulation of folding process in the presence of haem has been performed according to the procedure presented above, extended by the presence of haem in a simulation environment. Haem molecule (treated as five artificial amino-acid molecules) was free to translate in three directions and rotate around three axes (six degrees of freedom). The position of haem was dependent on its interaction with external force field (*fuzzy-oil-drop*) and on its interaction with folding polypeptide chain mutually influencing each other.

2.7 Structure analysis

Both the initial and final structures of haemoglobin, as well as intermediate structural forms (here the term ‘intermediate’ is used to describe the structure obtained after a particular step of folding simulation), were compared with the native structure of this protein according to the following criteria:

- The assessment of Secondary Structure (SS). SS was assigned to 3D structures using DSSP algorithm (Kabsch and Sander, 1983).
- The total number of Non-Bonding interactions (NB) assuming a cut-off distance of 12 Å.

- The Accessible Surface Area (ASA) taking a probe radius of 1.4 Å. ASA was calculated using surface racer programme (Tsodikov et al., 2002).
- The radius of gyration (R_g), which was calculated using the following equation (Flory, 1953):

$$R_g = \sqrt{\frac{1}{(N+1)^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left\langle\left(\vec{r}_i - \vec{r}_j\right)^2\right\rangle} \qquad (5)$$

  where N is the number of residues, $\vec{r}_i$ and $\vec{r}_j$ are the coordinates of Cα atom of ith and jth residue, respectively.
- The distances between the geometric centre of the molecule and sequential Cα atoms in the polypeptide chain ($D_{centre\text{-}C\alpha}$).
- RMSD-Cα calculation using the native form of haemoglobin as a reference structure. Structure alignments and RMSD calculations were done using VMD (Humphrey et al., 1996).
- The distances between haem iron atom and sequential Cα atoms of haemoglobin residues.
- The axis angles between adjoining helices. The axis angle is defined as the angle between the two helix vectors that correspond to the vector from the N-terminus atom to the C-terminus atom in two helices under consideration.

2.8 Hydrophobicity scale

The hydrophobicity scale is necessary to apply the presented model. Many scales for residue hydrophobicity are available. Some of them are based on analysis of known protein 3D structures (Kyte and Dooloittle, 1982; Eisenberg et al., 1982; Engelman et al., 1986; Hopp and Woods, 1981; Janin, 1979; Rose et al., 1985), while the others are derived from the physicochemical properties of amino-acid side chains (Wimley and White, 1996; Wolfender et al., 1981). The hydrophobicity scale including parameters for haem has been created using the relative position in *fuzzy-oil-drop* applied to single-domain proteins in PDB and particularly those being complexed with haem. The position found in this way was transformed into the hydrophobicity scale guaranteeing the unification of all amino acids and haem. Haem was divided into five fragments: haem_A, haem_B, haem_C, haem_D and haem_Fe (Figure 1). The calculation of hydrophobicity parameters was done utilising the set of 1353 structures of haem–protein complexes found in PDB No. 2004 (Berman et al., 2000). The complete hydrophobicity scale was calculated for all 20 amino acids and haem to make the scale consistent (Table 1).

Figure 1 Five virtual amino acids of haem: haem_A, haem_B, haem_C, haem_D and haem_Fe. For each virtual amino acids the hydrophobic parameter was assigned independently

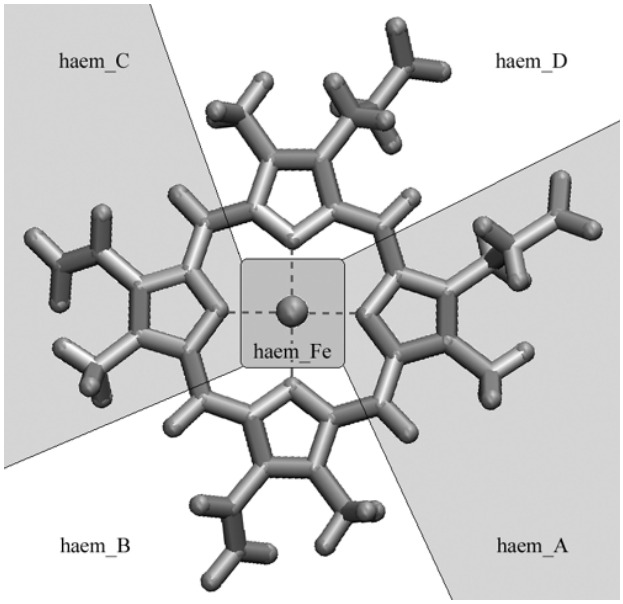


Table 1 Values of hydrophobicity parameter (H_i^r) assigned to each virtual residue distinguished in haem (italics) according to the *fuzzy-oil-drop* model, together with hydrophobicities of amino acids calculated elsewhere

Amino acid	*H-value*
LYS	0.000
GLU	0.083
ASP	0.167
GLN	0.250
ARG	0.272
ASN	0.278
PRO	0.300
SER	0.422
THR	0.478
haem_A	*0.532*
GLY	0.550
ALA	0.572
haem_D	*0.574*
HIS	0.628
TYR	0.700
haem_C	*0.716*
haem_Fe	*0.726*
haem_B	*0.763*

Table 1 Values of hydrophobicity parameter (H_i^r) assigned to each virtual residue distinguished in haem (italics) according to the *fuzzy-oil-drop* model, together with hydrophobicities of amino acids calculated elsewhere (continued)

Amino acid	*H-value*
LEU	0.783
VAL	0.811
MET	0.828
TRP	0.856
ILE	0.883
PHE	0.906
CYS	1.000

Source: Brylinski et al. (2006b)

3 Results

3.1 Active-site localisation in a protein molecule

The idealised distribution of hydrophobicity is represented by the 3-dimensional Gaussian function. The empirical distribution of hydrophobicity is expressed by the function describing residue–residue interaction introduced in Levitt (1976). The comparison between idealised hydrophobic distribution and the empirical one as observed in real proteins (crystal structure) reveals the discrepancies, which appeared to be biological-function dependent. The area of high difference points out the active site of a protein as well as the protein–protein contacts area in protein–protein complex creation. Figure 2 represents the results. The one-dimensional profiles of $\Delta\tilde{H}$ (left column) visualise also the colour scale applied. The same colour scale adopted for the 3-dimensional distribution of $\Delta\tilde{H}$ is shown in right column. The binding site distinguished by red and orange colour seems to be recognised very well in respect to ligand position (dark blue lines). Figure 3 shows also the localisation of the area critical for protein–protein complex creation. The $\Delta\tilde{H}$ profiles and 3-dimensional representations are shown as they appear in each protein – complex participant – separately. It is shown that both partners' binding sites can be easily recognised.

3.2 Simulation of protein folding

The conclusion based on the above analysis is that the native structure of protein represents the product of folding process of aim-oriented character. If so, the presence of external force field expressed by *fuzzy-oil-drop* may be used to direct the folding process towards hydrophobic core creation, which, modified by active presence of ligand (haem), can adopt the aim-oriented character. The simulation of protein-folding process *in silico* applying both assumptions is adopted to α and β chains of haemoglobin folding.

Figure 2 One-dimensional profiles of $\Delta\tilde{H}$ per amino acid (left column) and 3-dimensional distribution of $\Delta\tilde{H}$ on protein surface (right column) for (A) goose lysozyme complexed with NAG, (B) sperm whale myoglobin complexed with haem, (C) bovine acyl-coenzyme A complexed with palmitylo-coenzyme A, (D) catalytic domain of human TNF-α converting enzyme complexed with INN, (E) human lysozyme complexed with tetra-acetyl-chitotetraose, (F) bovine ribonuclease A complexed with 3',5'-DCPDG, (G) dihydrofolate reductase from *E. coli* complexed with methotrexate and (H) human mitogen-activated protein kinase complexed with SB203580. A colour scale is applied to expresses the magnitude of difference $\Delta\tilde{H}$ in particular protein surface area. The dark-blue (thick line representation) ligands are localised in their binding sites according to crystal structures

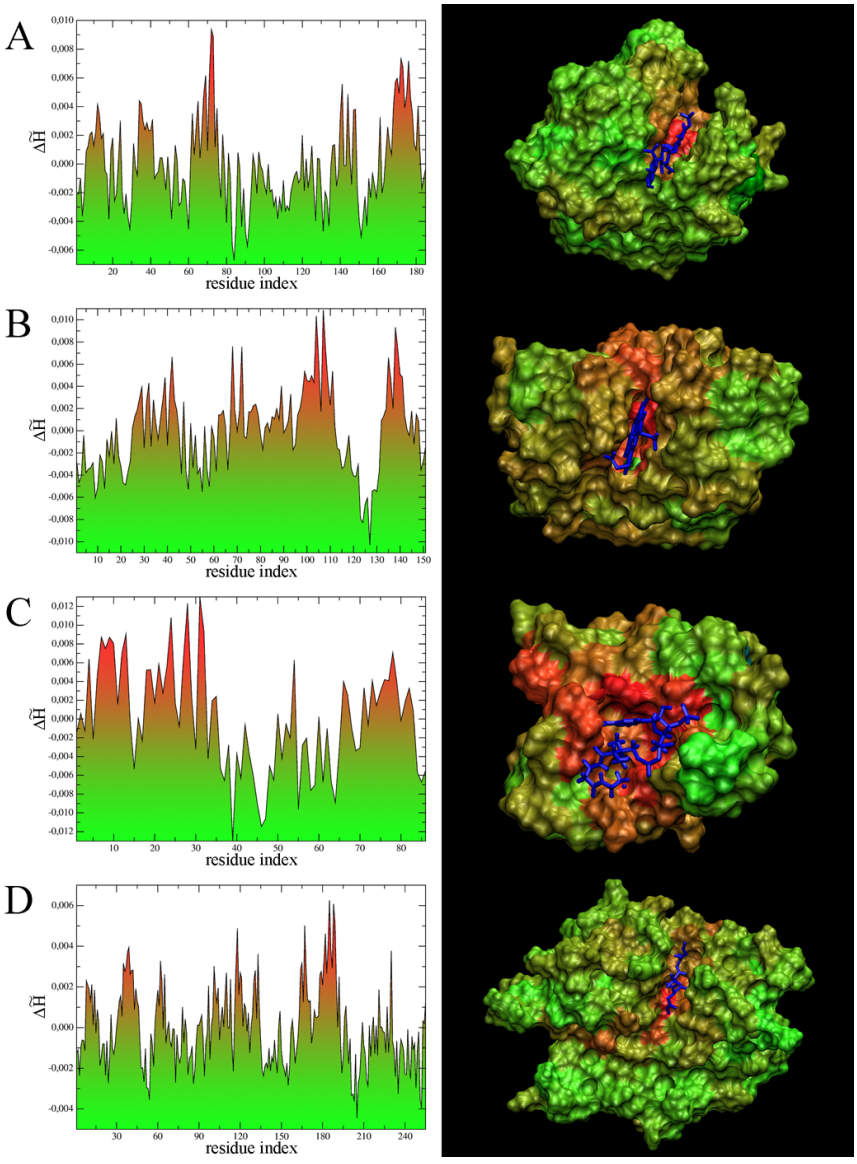


Figure 2 One-dimensional profiles of $\Delta\tilde{H}$ per amino acid (left column) and 3-dimensional distribution of $\Delta\tilde{H}$ on protein surface (right column) for (A) goose lysozyme complexed with NAG, (B) sperm whale myoglobin complexed with haem, (C) bovine acyl-coenzyme A complexed with palmitylo-coenzyme A, (D) catalytic domain of human TNF-α converting enzyme complexed with INN, (E) human lysozyme complexed with tetra-acetyl-chitotetraose, (F) bovine ribonuclease A complexed with 3',5'-DCPDG, (G) dihydrofolate reductase from *E. coli* complexed with methotrexate and (H) human mitogen-activated protein kinase complexed with SB203580. A colour scale is applied to expresses the magnitude of difference $\Delta\tilde{H}$ in particular protein surface area. The dark-blue (thick line representation) ligands are localised in their binding sites according to crystal structures (continued)

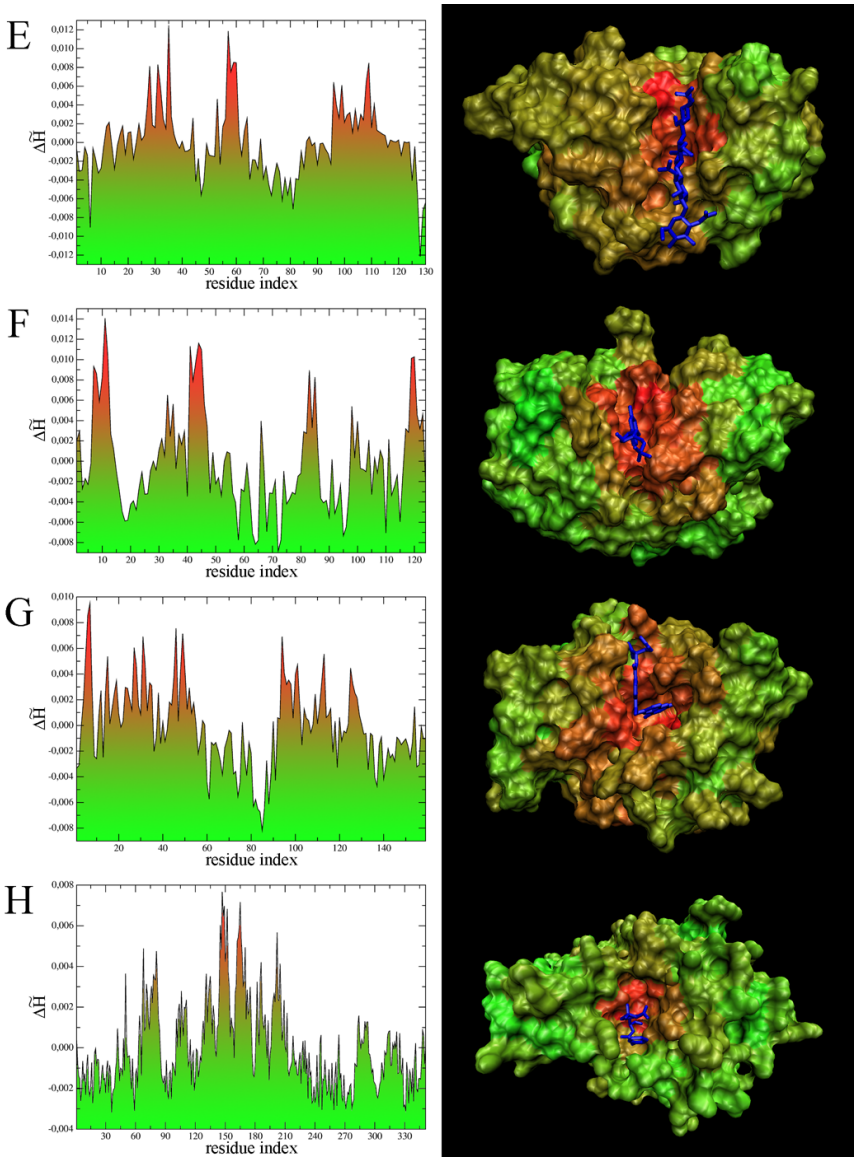
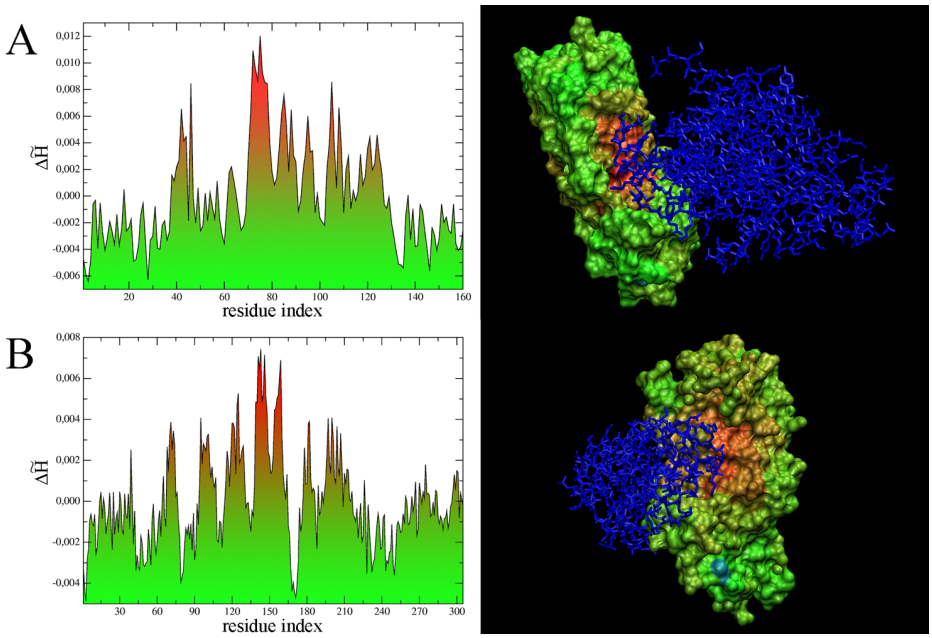


Figure 3 One-dimensional profiles of $\Delta\tilde{H}$ per amino acid (left column) and 3-dimensional distribution of $\Delta\tilde{H}$ on protein surface (right column) for (A) cell-cycle inhibitor P19INK4D (surface representation) complexed with human cyclin-dependent kinase CDK6 (line representation) and (B) inversely, CDK6 (surface representation) complexed with P19INK4D (line representation). A colour scale is applied to expresses the magnitude of difference $\Delta\tilde{H}$ in particular protein surface area



The simulation of folding process directed by external hydrophobic force field of *fuzzy-oil-drop* form is assumed to produce the protein molecule perfectly well soluble (low hydrophobic residues on the surface and high hydrophobic residues in a central part of molecule), with no tendency to interact with other molecules. The active presence of haem is assumed to direct the folding process towards the aim-oriented (ligand binding) folding, ensuring the ligand binding necessary for biological activity.

The appropriate parameterisation (hydrophobicity) of haem is necessary (the hydrophobicity scale is shown in Methods). The haem was free to move and rotate (six degrees of freedom), taking the appropriate orientation according to external force field and according to the surrounding residues of folding chain during the simulation.

3.2.1 Starting structures for folding process simulation

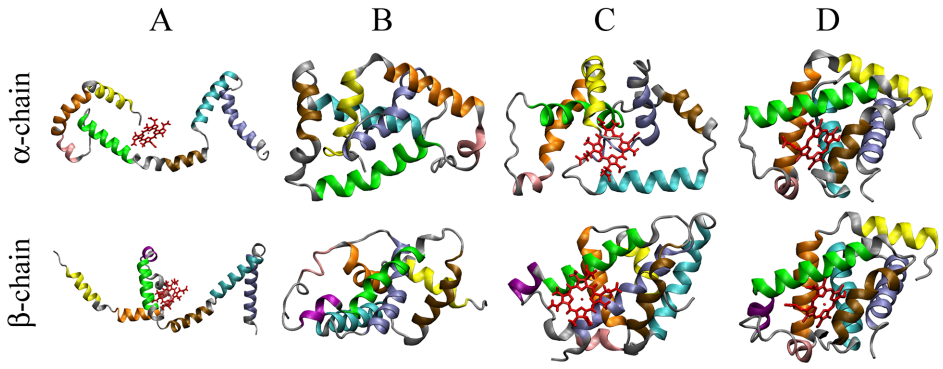
A large body of experiments and theoretical evidence suggests that local structure is frequently encoded in short segments of the protein sequence (Han and Baker, 1996; Fetrow et al., 1997; Bystroff and Baker, 1998; de Brevern et al., 2000, 2002). The method of protein structure prediction based on sequence-to-structure contingency table for tetrapeptides (Brylinski et al., 2005) allows to obtain early-stage structures for any amino-acid sequence (Brylinski et al., 2004a). The detailed description of early-stage folding model can be found in Jurkowski et al. (2004a) and Roterman (1995a, 1995b). The model applied to some well-known proteins positively verified the limited conformational sub-space (Jurkowski et al., 2004a, 2004b; Brylinski et al. 2004a, 2004b).

The recently published results of molecular dynamics simulation of 3- and 21-alanine polypeptide in the temperature range 276–402 K additionally positively verified the ellipse path as the path along which the helix unfolding is taking place (Paschek et al., 2005; Gnanakaran and Garcia, 2005).

The global analysis of all available proteins produced the sequence-to-structure contingency table, which can be used as a standard tool for early-structure prediction (Brylinski et al., 2004c, 2005). This contingency table has been used to create the early-stage structure for haemoglobin chains (Brylinski et al., 2004b).

The sequence of human haemoglobin was used as input for early-stage structure prediction. The Structure Predictability Index (SPI) (Brylinski et al., 2004a) calculated for the sequence of α and β chains was found to be 99.2 and 98.9, respectively. Those values rank haemoglobin as a very easy target for early-stage form prediction. The Q3 (Rost and Sander, 1993), Q7 (Brylinski et al., 2004a) and SOV (Rost et al., 1994; Zemla et al., 1999) parameters calculated vs. native structure were found to be 93.4, 94.3 and 86.0 for α chain and 93.1, 94.0 and 85.8 for β chain, respectively. The extremely high accuracy of early-stage structure prediction for haemoglobin justified the selection of this form as the good starting structure for late-stage folding simulation. The prediction, including SPI, Q3 and Q7 estimations, can be easily carried out for any protein sequence with free prediction server available from http://bioinformatics.cm-uj.krakow.pl/earlystage. The starting structures of both chains are shown in Figure 4.

Figure 4 The structures of both haemoglobin chains as they appeared in early-stage folding step (A), both final forms: with haem molecule absent (B) and present (C) in simulations. For comparison, native structures are shown in (D). The structures are coloured according to secondary structure assignment for the native structures: helix A (yellow), B (orange), C (pink), D (purple), E (green), F (ochre), G (cyan) and H (ice blue). Haem molecule is shown in red



3.2.2 Monitoring of the folding process

The structure of α and β haemoglobin chain is highly helical. The mutual spatial arrangement of the helices is very representative for these molecules. The angles between axes of adjoining helices were monitored during folding simulation in absence and presence of haem. The results are shown in Figure 5. The presence of haem seems to help to reach proper orientation of following helices in α chain of haemoglobin: B–C, E–F and G–H and in β chain of haemoglobin: B–C, C–D, D–E and F–G. Significantly worst results were found in the presence of haem in α chain of haemoglobin A–B and

β chain of haemoglobin A–B, E–F and G–H. The A–B orientation is difficult to be interpreted. It seems that their improper orientation does not depend on the presence of haem.

The optimisation procedure applied in folding simulation is an iterative procedure in which each hydrophobic interaction-driven step is followed by energy minimisation step. The fluctuations of parameters monitored during folding process for these two approaches are shown in Figure 6. Hydrophobic effects, which cause non-polar side chains to tend to cluster together in the protein interior, can be measured by the solvent ASA. The radius of gyration (R_g) and Non-Bonding interactions (NB) estimate the characteristic volume of the globular protein and provides quantitative information on its compactness.

Figure 5 The change of angles between axes of adjoining helices during folding simulation, with haem absent (grey lines) and present (black lines) in folding process. The constant dotted line points at the value of particular axis angle as appears in native structure

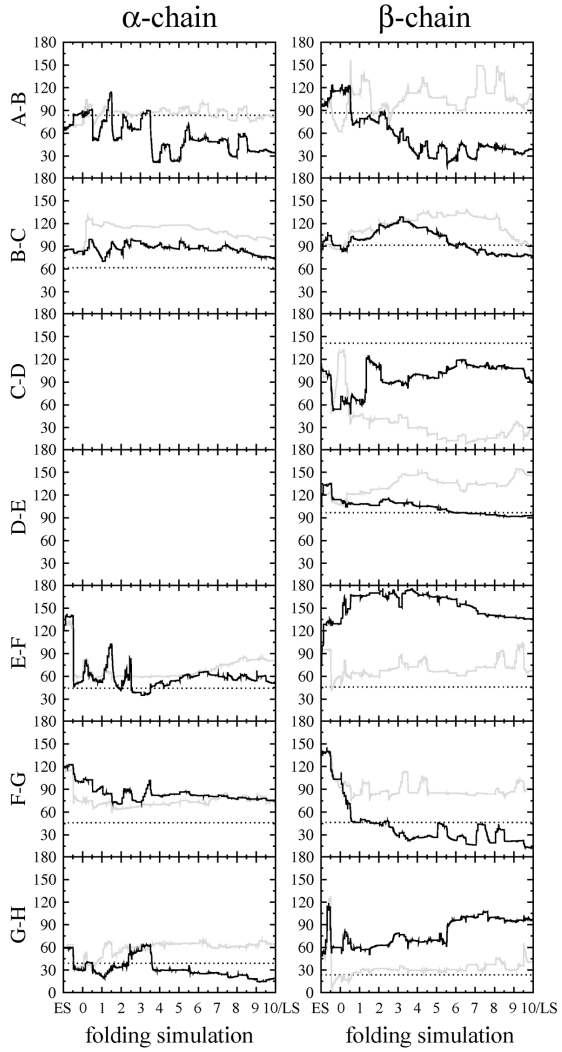
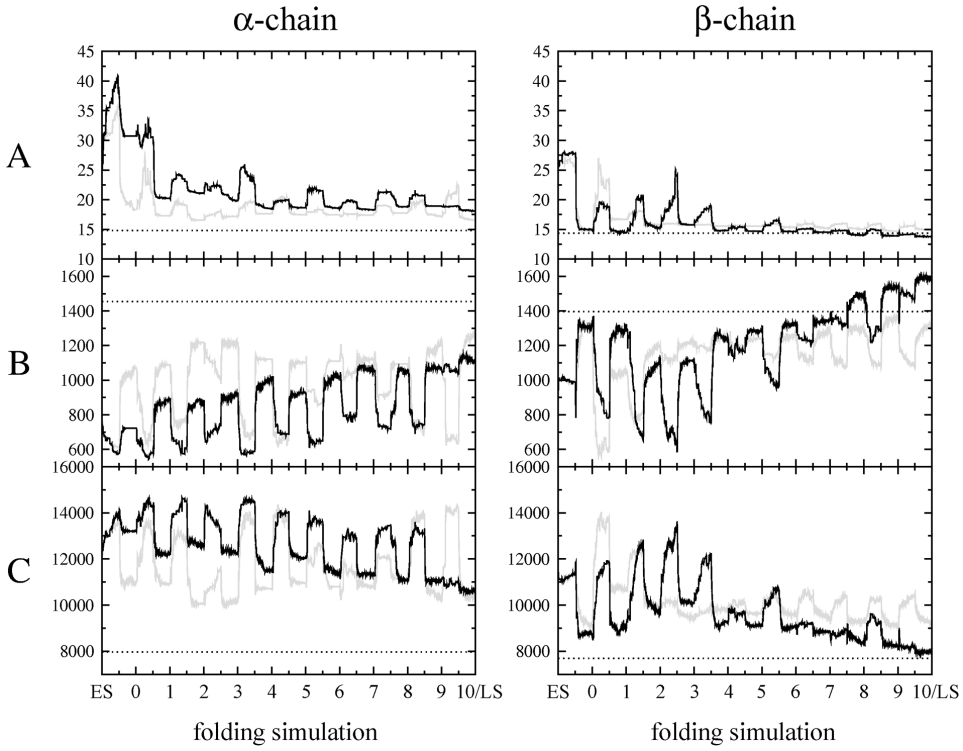


Figure 6 The time profiles of parameters monitored during folding process: the radius of gyration [Å] (A), the total number of non-bonding interactions (B) and solvent-accessible surface area (ASA) [Å^2] (C). The simulations with haem molecule absent and present in folding process are represented by grey and black lines, respectively. In all plots, the constant dotted line pinpoints the quantity of a probed parameter (R_g, NB or ASA) calculated for native structure. ES denotes early stage treated as starting one for simulation and LS – late stage treated as final one for simulation



3.2.3 The non-bonding contacts and RMSD calculation

The comparison of non-bonding contact maps is used commonly to verify the correctness of obtained structures. These maps for all discussed structural forms are shown in Figure 7. Table 2 presents also the percentage of native non-bonding contacts in final structures obtained according to the adopted model, together with RMSD-Cα using the native forms as reference structures. Noticeable better results are seen for β chain of haemoglobin.

Figure 7 Non-bonding contact maps for haemoglobin: early-stage of folding (A), folded *in silico* in the absence (B) and presence (C) of haem. For comparison, non-bonding contact maps for native form of haemoglobin are shown in (D)

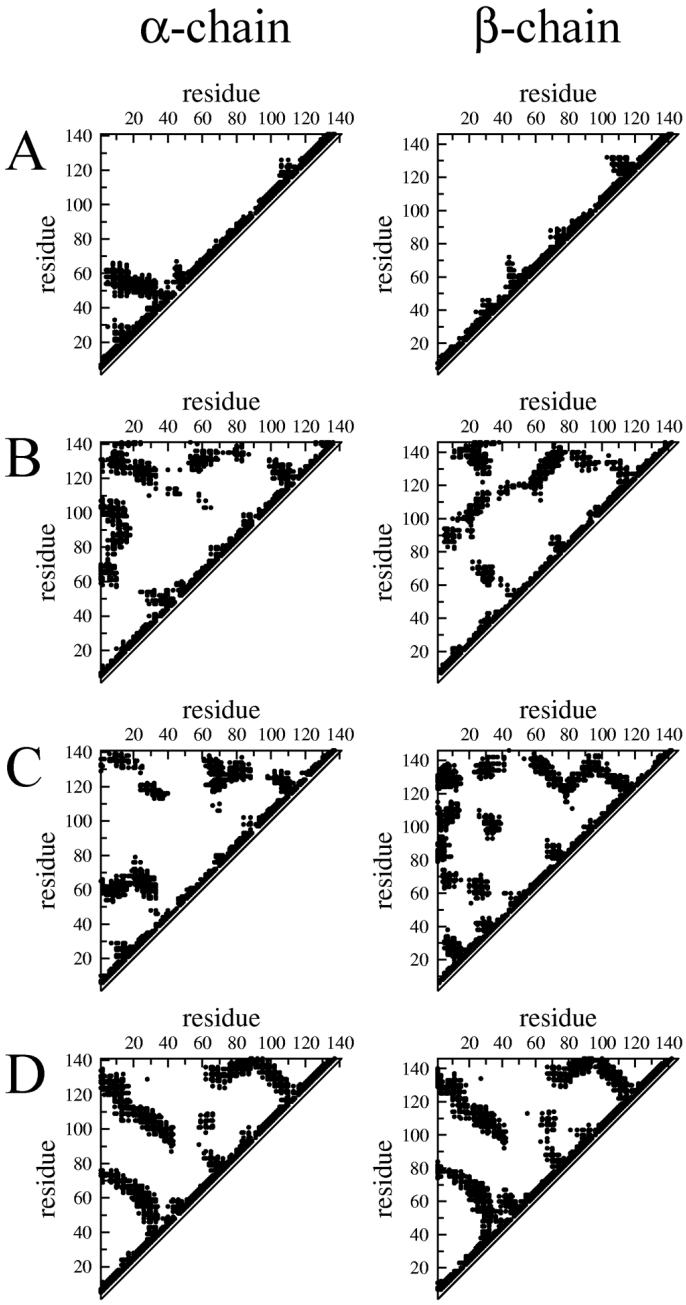


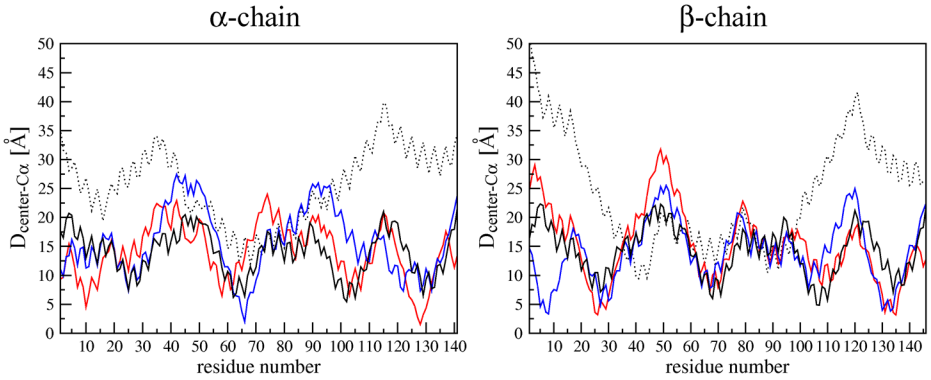
Table 2 The percentage of native non-bonding contacts (NB_{nat}) and RMSD-Cα for all discussed structures of haemoglobin

Chain	*Structure*	*NB_{nat} (%)*	*RMSD-Cα*
α	Early-stage of folding *in silico*	40.37	22.49
	Folded *in silico* in the absence of haem	37.80	14.24
	Folded *in silico* in the presence of haem	39.80	14.83
β	Early-stage of folding *in silico*	37.69	25.02
	Folded *in silico* in the absence of haem	36.86	15.82
	Folded *in silico* in the presence of haem	40.51	10.15

3.2.4 Spatial distribution of Cα atoms vs. the geometrical centre

The profile of the length of vectors linking geometrical centre of the molecule with Cα atoms of sequential amino acids are shown in Figure 8. The best results in this category of classification seems to be received for β chain of haemoglobin, although the α chain of haemoglobin also seems to represent relatively high accordance. In both the cases, the presence of haem during the folding simulation improved the spatial organisation of a protein.

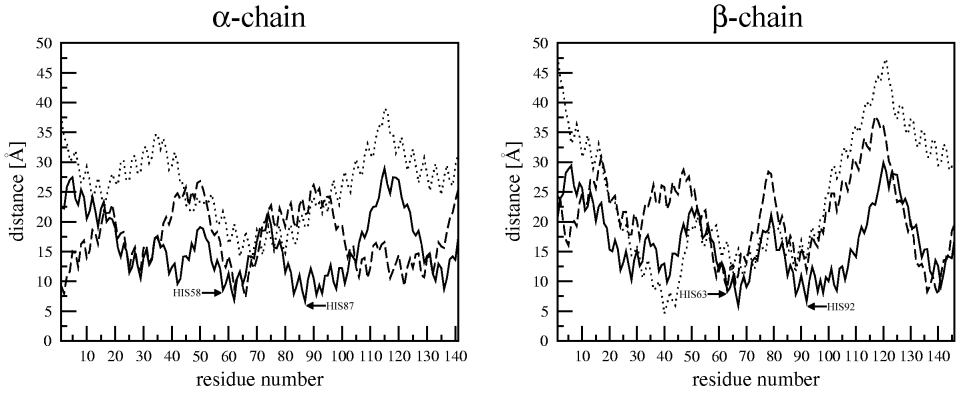
Figure 8 Profiles of $D_{centre\text{-}C\alpha}$ vectors for early stage of folding (black dotted lines), folded *in silico* in the absence (blue solid lines) and presence (red solid lines) of haem, and native (black solid lines) structure of haemoglobin



3.2.5 Haem-binding site

The critical for biological activity of haem-binding proteins is the presence of the ligand and its orientation vs. the host protein molecule. The His58-distal and His87-proximal histidines are responsible for haem binding in α chain of haemoglobin. The distances between haem iron atom and sequential Cα atoms of haemoglobin residues are shown in Figure 9. His58 seems to reach a proper localisation vs. haem, although His87 does not. The haem-binding histidines in β chain of haemoglobin are His63-distal and His92-proximal. In the structure of β chain of haemoglobin obtained in the presence of haem during folding process, both histidines are localised in the close vicinity of haem. The profiles of distances between iron atom of haem and sequential Cα for the native and folded *in silico* structures of β chain are in remarkably good agreement.

Figure 9 The distances between haem iron atom and sequential Cα atoms of haemoglobin residues for early stage of folding (dotted line), folded *in silico* in the presence of haem (dashed line), and native (solid line) structure



4 Discussion

4.1 Active-site localisation

The model based on *fuzzy-oil-drop* seems to work well as the marker for active-site localisation. The contact areas in proteins creating complexes also seem to be recognised successfully. Method is very simple. It can be particularly useful in recognition of a large number of proteins synthesised on the basis of genome analysis of unknown biological function, which wait for unified automated method for their biological activity recognition (Burley et al., 1999). The method is also easily applicable to fully automated system. The large-scale calculation can verify the model as the universal one. This type of work has already been undertaken.

The correctly localised binding cavity makes possible the construction of the compatible ligand (its shape and chemical characteristics), what can be performed using classical *de novo* design methods for ligand construction (Baurin et al., 2000; Cramer et al., 1989; Polanski and Walczak, 2000; Sippl, 2002).

4.2 Simulation of protein-folding process rather than protein structure prediction

The X-ray structure deposited in PDB represents the form of α–β dimer (Fermi et al., 1984). The α and β chains of haemoglobin represent very similar structure, although their folding process differs (Baglioni, 1968; Bunn, 1987). The significant lower stability of α chain has been experimentally verified (Feng et al., 2004, 2005; Scott et al., 1993). Its structural instability may cause the improper exposure of haem group, allowing the participation of iron ion in unexpected redox processes (Feng et al., 2004; Scott et al. 1993). The protective mechanism of AHSP aimed at proper haem stabilisation in an α-globin polypeptide chain has been discussed recently (Feng et al., 2005). It suggests that the α chain of haemoglobin needs external support in form of AHSP polypeptide, which when bound ensures proper structural form mostly oriented on correct haem binding, although only by distal His58 (Feng et al., 2004; Scott et al., 1993).

The differences between α and β chains observed in simulated folding process chains seem to be in accordance with experimental observations. The difficulties in α-chain structure prediction seem to be originated in its natural biological character. The applied procedure has not produced the native structure of haemoglobin. This process seems to be of hierarchy nature, taking into account step-wise folding, in which α chain possibly folds in the company of more stable β chain. The stabilisation role of AHSP as well as possible mutual influence between α and β chain interaction on the folding process will be taken into consideration in further simulations with *fuzzy-oil-drop* model.

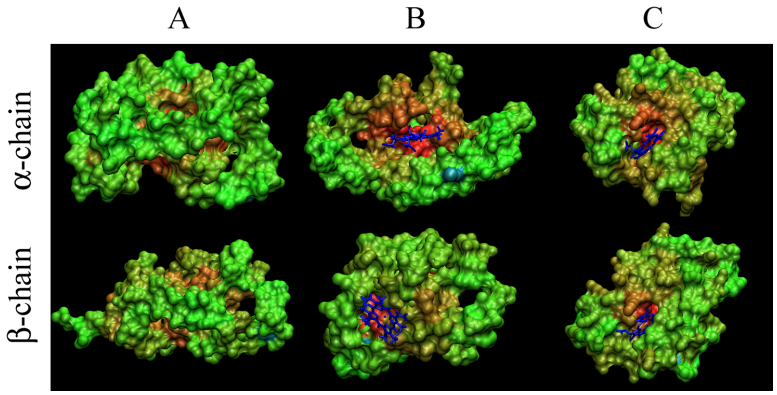
The idea of external hydrophobic force field introduced here seems to be extended taking also other environment-originated elements. The recently described mechanism of protein folding in the presence of trigger factor, which interacts with freshly synthesised polypeptide chain leaving the channel of ribosome (Ferbitz et al., 2004), seems also to interpret the folding process as directed one. The same role of external force field can be played by chaperonins, the interior of which represents hydrophilic character, which guaranties exposure of hydrophilic residues towards the protein surface (Braig et al., 1994; Houry et al., 1999; Wang et al., 1999). The *fuzzy-oil-drop* model takes into account the influence of polar environment, pushing the hydrophobic residues to occupy the position rather in a centre of the molecule and hydrophilic ones on the surface of molecule. The chaperonin simultaneously obliges the protein molecule to be squeezed, what the presented model of *fuzzy-oil-drop* is also able to do, playing the role of controlled hydrophobicity-driven implosion.

Structural forms of haemoglobin were received as the result of simulated folding process in the presence of external force field assumed to represent the process driven by hydrophobic interactions. When proteins were diluted from concentrated denaturant at the beginning of a stopped-flow folding experiment, they very generally exhibited a fast chain contraction on a submillisecond time scale. This 'burst phase' behaviour has often been interpreted in terms of the fast formation of productive folding intermediates (Roder and Colon, 1997). The experimental results for ribonuclease A (Qi et al., 1998) and analogous results for cytochrome *c* (Sosnick et al., 1996, 1997) indicated that the submillisecond burst phase seen for these proteins reflects a solvent-dependent readjustment of the unfolded state ensemble, rather than the rapid formation of distinct folding intermediates. Those experimental results, together with theoretical studies on protein folding (Sosnick et al., 2002; Fernandez et al., 2002), suggest that the initial step in globular protein folding is a long-range conformational search to find a topologically native-like transition-state nucleus. The rapid compaction of the polypeptide chain occurring at the beginning of the folding process was found to be driven by strong hydrophobic interactions (Nozaki and Tanford, 1970; Gutin et al. 1995) and buried a great deal of surface (Sosnick et al., 1997). The results obtained for hydrophobic-interaction-driven steps correspond well with the 'burst phase' observed in folding experiments. In response to the new conditions occurring at each squeezing step, the rapid compaction of the polypeptide chain was observed. As it might be expected, the presence of haem did not change the general characteristics of folding process. Similar folding route was observed during ribonuclease A-simulated folding, according to presented *fuzzy-oil-drop* model (Brylinski et al., 2004b, 2005, 2006b, 2006c).

4.3 Aim orientation of protein folding

The question arises whether the biological function can be present in molecules received as a result of simulation. The biological activity was assumed to be identified as aim-oriented discrepancy between idealised and observed hydrophobic oil drop (Brylinski et al., 2006c). The comparison of the native structure of haemoglobin (α and β chain) with both structures received, according to folding procedure presented in this paper, reveals the aim orientation of folding process. Figure 10 shows the distribution of hydrophobicity irregularities $\Delta\tilde{H}$ vs. the idealised *fuzzy-oil-drop*. The main, very easily recognised difference is that, in the native structure of protein, the highest irregularity vs. the idealised *fuzzy-oil-drop* is localised in the active centre. The structures received in the absence of haem in folding procedure, characterised by more regular hydrophobicity distribution, with hydrophilic residues exposed on the surface as assumed in a *fuzzy-oil-drop* model, probably exemplify inactive proteins. The results obtained for both chains of haemoglobin show that the presence of natural ligand in folding process seems to be important in simulation of protein-folding process.

Figure 10 The localisation of hydrophobicity irregularities $\Delta\tilde{H}$ vs. the idealised *fuzzy-oil-drop* for structures folded *in silico* in the absence (A) and presence (B) of haem, together with native (C) structure of haemoglobin. The continuous colour scale is applied to express low (green) to high (red) values of $\Delta\tilde{H}$. Haem molecule is shown as dark blue



Acknowledgements

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 WŁ/222/P/L.

References

Aguilar, C.F., Thomas, P.J., Mills, A., Moss, D.S. and Palmer, R.A. (1992) 'Newly observed binding mode in pancreatic ribonuclease', *J. Mol. Biol.*, Vol. 224, pp.265–267.

Baglioni, C. (1968) 'Chromosomal and cytoplasmic regulation of haemoglobin synthesis', *Bibl. Haematol.*, Vol. 29, pp.1056–1063.

Baldwin, R.L. (2002) 'Making a network of hydrophobic clusters', *Science*, Vol. 295, pp.1657–1658.

Baumann, G., Frommel, C. and Sander, C. (1989) 'Polarity as a criterion in protein design', *Protein Eng*., Vol. 2, pp.329–334.

Baurin, N., Vangrevelinghe, E., Morin-Allory, L., Merour, J.Y., Renard, P., Payard, M., Guillaumet, G. and Marot, C. (2000) '3D-QSAR CoMFA study on imidazolinergic I(2) ligands: a significant model through a combined exploration of structural diversity and methodology', *J. Med. Chem*., Vol. 43, pp.1109–1122.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) 'The protein data bank', *Nucleic Acids Res*., Vol. 28, pp.235–242.

Bonneau, R., Strauss, C.E. and Baker, D. (2001) 'Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation', *Proteins*, Vol. 43, pp.1–11.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (1998) 'Predicting function: from genes to genomes and back', *J. Mol. Biol*., Vol. 283, pp.707–725.

Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D.C., Joachimiak, A., Horwich, A.L. and Sigler, P.B. (1994) 'The crystal structure of the bacterial chaperonin GroEL at 2.8 A', *Nature*, Vol. 371, pp.578–586.

Brotherton, D.H., Dhanaraj, V., Wick, S., Brizuela, L., Domaille, P.J., Volyanik, E., Xu, X., Parisini, E., Smith, B.O., Archer, S.J., Serrano, M., Brenner, S.L., Blundell, T.L. and Laue, E.D. (1998) 'Crystal structure of the complex of the cyclin D-dependent kinase Cdk6 bound to the cell-cycle inhibitor p19INK4d', *Nature*, Vol. 395, pp.244–250.

Brylinski, M., Konieczny, L. and Roterman, I. (2004a) 'SPI – structure predictability index for protein sequences', *In Silico Biology*, Vol. 5, p.22.

Brylinski, M., Jurkowski, W., Konieczny, L. and Roterman, I. (2004b) 'Limited conformational space for early-stage protein folding simulation', *Bioinformatics*, Vol. 20, pp.199–205.

Brylinski, M., Jurkowski, W., Konieczny, L. and Roterman, I. (2004c) 'Limitation of conformational space for proteins – early stage folding simulation of human α and β hemoglobin chains', *Task Quarterly*, Vol. 8, pp.413–422.

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.*, Vol. 2, pp.65–79.

Brylinski, M., Konieczny, L. and Roterman, I. (2006a) *Ligation Site in Proteins Recognized in Silico Bioinformation*, Vol. 1, No. 4, pp.127–129.

Brylinski, M., Konieczny, L. and Roterman, I. (2006b) 'Fuzzy-oil-drop hydrophobic force field – a model to represent late-stage folding (In Silico) of lysozyme', *J. Biomol. Struct. Dyn*., Vol. 23, pp.519–528.

Brylinski, M., Kochanczyk, M., Broniatowska, E. and Roterman, I. (2007) 'Localization of ligand binding site in proteins identified in silico', *J. Mol. Model*., PMID: 17394030.

Brylinski, M., Konieczny, L. and Roterman, I. (2006d) 'Hydrophobic collapse in late-stage folding (*in silico*) of bovine pancreatic trypsin inhibitor', *Biochimie*, Vol. 88, pp.1229–1239.

Bunn, H.F. (1987) 'Subunit assembly of hemoglobin: an important determinant of hematologic phenotype', *Blood*, Vol. 69, pp.1–6.

Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W. and Swaminathan, S. (1999) 'Structural genomics: beyond the human genome project', *Nat. Genet*., Vol. 23, pp.151–157.

Bystroff, C. and Baker, D. (1998) 'Prediction of local structure in proteins using a library of sequence-structure motifs', *J. Mol. Biol*., Vol. 281, pp.565–577.

Cramer III., R.D., Patterson, D.E. and Bunce, J.D. (1989) 'Recent advances in comparative molecular field analysis (CoMFA)', *Prog. Clin. Biol. Res.*, Vol. 291, pp.161–165.

Crippen, G.M. and Kuntz, I.D. (1978) 'A survey of atom packing in globular proteins', *Int. J. Pept. Protein Res*., Vol. 12, pp.47–56.

de Brevern, A.G., Etchebest, C. and Hazout, S. (2000) 'Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks', *Proteins*, Vol. 41, pp.271–287.

de Brevern, A.G., Valadie, H., Hazout, S. and Etchebest, C. (2002) 'Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship', *Protein Sci*., Vol. 11, pp.2871–2886.

Devos, D. and Valencia, A. (2000) 'Practical limits of function prediction', *Proteins*, Vol. 41, pp.98–107.

Dill, K.A. (1990) 'Dominant forces in protein folding', *Biochemistry*, Vol. 29 pp.7133–7155.

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*., Vol. 82, pp.2609–2616.

Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) 'Analysis of membrane and surface protein sequences with the hydrophobic moment plot', *J. Mol. Biol*., Vol. 179, pp.125–142.

Eisenberg, D., Weiss, R.M., Terwilliger, T.C. and Wilcox, W. (1982) 'Hydrophobic moments of protein structure', *Faraday Symp. Chem. Soc.*, Vol. 17, pp.109–120.

Elcock, A.H. (2001) 'Prediction of functionally important residues based solely on the computed energetics of protein structure', *J. Mol. Biol.*, Vol. 312, pp.885–896.

Engelman, D.M. and Zaccai, G. (1980) 'Bacteriorhodopsin is an inside-out protein', *Proc. Natl. Acad. Sci. USA*, Vol. 77, pp.5894–5898.

Engelman, D.M., Steitz, T.A. and Goldman, A. (1986) 'Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins', *Annu. Rev. Biophys. Biophys. Chem*., Vol. 15, pp.321–353.

Feng, L., Gell, D.A., Zhou, S., Gu, L., Kong, Y., Li, J., Hu, M., Yan, N., Lee, C., Rich, A.M., Armstrong, R.S., Lay, P.A., Gow, A.J., Weiss, M.J., Mackay, J.P. and Shi, Y. (2004) 'Molecular mechanism of AHSP-mediated stabilization of alpha-hemoglobin', *Cell*, Vol. 119, pp.629–640.

Feng, L., Zhou, S., Gu, L., Gell, D.A., Mackay, J.P., Weiss, M.J., Gow, A.J. and Shi, Y. (2005) 'Structure of oxidized alpha-haemoglobin bound to AHSP reveals a protective mechanism for haem', *Nature*, Vol. 435, pp.697–701.

Ferbitz, L., Maier, T., Patzelt, H., Bukau, B., Deuerling, E. and Ban, N. (2004) 'Trigger factor in complex with the ribosome forms a molecular cradle for nascent proteins', *Nature*, Vol. 431, pp.590–596.

Fermi, G., Perutz, M.F., Shaanan, B. and Fourme, R. (1984) 'The crystal structure of human deoxyhaemoglobin at 1.74 A resolution', *J. Mol. Biol*., Vol. 175, pp.159–174.

Fernandez, A., Colubri, A. and Berry, R.S. (2002) 'Three-body correlations in protein folding: the origin of cooperativity', *Physica A*, Vol. 307, pp.235–259.

Fetrow, J.S., Palumbo, M.J. and Berg, G. (1997) 'Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme', *Proteins*, Vol. 27, pp.249–271.

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*., Vol. 105, pp.391–409.

Flory, P.J. (1953) *Principles of Polymer Chemistry*, Cornell University Press, Ithaca.

Gnanakaran, S. and Garcia, A.E. (2005) 'Helix-coil transition of alanine peptides in water: force field dependence on the folded and unfolded structures', *Proteins*, Vol. 59, pp.773–782.

Gutin, A.M., Abkevich, V.I. and Shakhnovich, E.I. (1995) 'Evolution-like selection of fast-folding model proteins', *Proc. Natl. Acad. Sci. USA*, Vol. 92, pp.1282–1286.

Han, K.F. and Baker, D. (1996) 'Global properties of the mapping between local amino acid sequence and local structure in proteins', *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp.5814–5818.

Hegyi, H. and Gerstein, M. (1999) 'The relationship between protein structure and function: a comprehensive survey with application to the yeast genome', *J. Mol. Biol.*, Vol. 288, pp.144–164.

Holm, H. and Sander, C. (1992) 'Evaluation of protein models by atomic solvation preference', *J. Mol. Biol*., Vol. 225, pp.93–105.

Hopp, T.P. and Woods, K.R. (1981) 'Prediction of protein antigenic determinants from amino acid sequences', *Proc. Natl. Acad. Sci. USA*, Vol. 78, pp.3824–3828.

Houry, W.A., Frishman, D., Eckerskorn, C., Lottspeich, F. and Hartl, F.U. (1999) 'Identification of in vivo substrates of the chaperonin GroEL', *Nature*, Vol. 402, pp.147–154.

Humphrey, W., Dalke, A. and Schulten, K. (1996) 'VMD: visual molecular dynamics', *J. Mol. Graph*, Vol. 14, pp.27–38.

Izrailev, S. and Farnum, M.A. (2004) 'Enzyme classification by ligand binding', *Proteins*, Vol. 57, pp.711–724.

Janin, J. (1979) 'Surface and inside volumes in globular proteins', *Nature*, Vol. 277, pp.491–492.

Janin, J. (2005a) 'The targets of CAPRI rounds 3-5', *Proteins*, Vol. 60, pp.170–175.

Janin J. (2005b) 'Assessing predictions of protein-protein interaction: the CAPRI experiment', *Protein Sci*., Vol. 14, pp.278–283.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I. and Wodak, S.J. (2003) 'CAPRI: a critical assessment of predicted interactions', *Proteins*, Vol. 52, pp.2–9.

Jimenez, J.L. (2005) 'Does structural and chemical divergence play a role in precluding undesirable protein interactions?', *Proteins*, Vol. 59 pp.757–764.

Jones, S. and Thornton, J.M. (1997) 'Analysis of protein-protein interaction sites using surface patches'. *J. Mol. Biol*., Vol. 272, pp.121–132.

Jurkowski, W., Brylinski, M., Konieczny, L. and Roterman, I. (2004a) 'Lysozyme folded in silico according to the limited conformational sub-space', *J. Biomol. Struct. Dyn*., Vol. 22, pp.149–158.

Jurkowski, W., Brylinski, M., Konieczny, L., Wiiniowski, Z. and Roterman, I. (2004b) 'Conformational subspace in simulation of early-stage protein folding', *Proteins*, Vol. 55, pp.115–127.

Kabsch, W. and Sander, C. (1983) 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers*, Vol. 22, pp.2577–2637.

Kanaya, S., Oobatake, M. and Liu, Y. (1996) 'Thermal stability of Escherichia coli ribonuclease HI and its active site mutants in the presence and absence of the Mg2+ ion. Proposal of a novel catalytic role for Glu48', *J. Biol. Chem*., Vol. 271, pp.32729–32736.

Kauzmann, W. (1959) 'Some factors in the interpretation of protein denaturation', *Adv. Protein Chem*., Vol. 14 pp.1–63.

Klapper, M.H. (1971) 'On the nature of the protein interior', *Biochim. Biophys. Acta*, Vol. 229, pp.557–566.

Klotz, I.M. (1970) 'Comparison of molecular structures of proteins: helix content; distribution of apolar residues', *Arch. Biochem. Biophys*., Vol. 138 pp.704–706.

Konieczny, L., Brylinski, M. and Roterman, I. (2006) 'Gauss-function-based model of hydrophobicity density in proteins', *In Silico. Biology*, Vol. 6, pp.15–22.

Kragelund, B.B., Andersen, K.V., Madsen, J.C., Knudsen, J. and Poulsen, F.M. (1993) 'Three-dimensional structure of the complex between acyl-coenzyme A binding protein and palmitoyl-coenzyme A', *J. Mol. Biol*., Vol. 230, pp.1260–1277.

Kuntz, I.D. and Crippen, G.M. (1979) 'Protein densities' *Int. J. Pept. Protein Res*., Vol. 13, pp.223–228.

Kyte, J. and Doolittle, R.F. (1982) 'A simple method for displaying the hydropathic character of a protein', *J. Mol. Biol.*, Vol. 157 pp.105–132.

Lamb, M.L., Burdick, K.W., Toba, S., Young, M.M., Skillman, A.G., Zou, X., Arnold, J.R. and Kuntz, I.D. (2001) 'Design, docking, and evaluation of multiple libraries against multiple targets', *Proteins*, Vol. 42, pp.296–318.

Levitt, M. (1976) 'A simplified representation of protein conformations for rapid simulation of protein folding', *J. Mol. Biol.*, Vol. 104, pp.59–107.

Liang, J., Edelsbrunner, H. and Woodward, C. (1998) 'Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design', *Protein Sci*., Vol. 7, pp.1884–1897.

Luque, I. and Freire, E. (2000) 'Structural stability of binding sites: consequences for binding affinity and allosteric effects', *Proteins*, Suppl. 4, pp.63–71.

Maskos, K., Fernandez-Catalan, C., Huber, R., Bourenkov, G.P., Bartunik, H., Ellestad, G.A., Reddy, P., Wolfson, M.F., Rauch, C.T., Castner, B.J., Davis, R., Clarke, H.R., Petersen, M., Fitzner, J.N., Cerretti, D.P., March, C.J., Paxton, R.J., Black, R.A. and Bode, W. (1998) 'Crystal structure of the catalytic domain of human tumor necrosis factor-alpha-converting enzyme', *Proc. Natl. Acad. Sci. USA*, Vol. 95, pp.3408–3412.

Meiering, E.M., Serrano, L. and Fersht, A.R. (1992) 'Effect of active site residues in barnase on activity and stability', *J. Mol. Biol*., Vol. 225, pp.585–589.

Meirovitch, H. and Scheraga, H.A. (1980a) '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*, Vol. 13, pp.1406–1414.

Meirovitch, H. and Scheraga, H.A. (1980b) 'Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids', *Macromolecules*, Vol. 13, pp.1398–1405.

Meirovitch, H. and Scheraga, H.A. (1981) 'Empirical studies of hydrophobicity. 3. Radial distribution of clusters of hydrophobic and hydrophilic amino acids', *Macromolecules*, Vol. 14, pp.340–345.

Mendez, R., Leplae, R., de Maria, L. and Wodak, S.J. (2003) 'Assessment of blind predictions of protein–protein interactions: current status of docking methods', *Proteins*, Vol. 52, pp.51–67.

Mendez, R., Leplae, R., Lensink, M.F. and Wodak, S.J. (2005) 'Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures', *Proteins*, Vol. 60, pp.150–169.

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 occurring amino acids', *J. Phys. Chem*., Vol. 79, pp.2361–2381.

Nemethy, G., Pottle, M.S. and Scheraga, H.A. (1983) 'Energy parameters in polypeptides. 9. Updating of geometrical parameters, non-bonding interactions and hydrogen bonding interactions for naturally occurring amino acids', *J. Phys. Chem*., Vol. 87, pp.1883–1887.

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 non-bonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides', *J. Phys. Chem*., Vol. 96, pp.6472–6484.

Novotny, J., Rashin, A.A. and Bruccoleri, R.E. (1988) 'Criteria that discriminate between native proteins and incorrectly folded models', *Proteins*, Vol. 4, pp.19–30.

Nozaki, Y. and Tanford, C. (1970) 'The solubility of amino acids, diglycine, and triglycine in aqueous guanidine hydrochloride solutions', *J. Biol. Chem*., Vol. 245, pp.1648–1652.

Ondrechen, M.J., Clifton, J.G. and Ringe D. (2001) 'THEMATICS: a simple computational predictor of enzyme function from structure', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp.12473–12478.

Oshiro, C.M., Kuntz, I.D. and Dixon, J.S. (1995) 'Flexible ligand docking using a genetic algorithm', *J. Comput. Aided Mol. Des.*, Vol. 9 pp.113–130.

Paschek, D., Gnanakaran, S. and Garcia, A.E. (2005) 'Simulations of the pressure and temperature unfolding of an alpha-helical peptide', *Proc. Natl. Acad. Sci. USA*, Vol. 102, No. 19, pp.6765–6770.

Polanski, J. and Walczak, B. (2000) 'The comparative molecular surface analysis (COMSA): a novel tool for molecular design', *Comput. Chem*., Vol. 24, pp.615–625.

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*., Vol. 5, pp.882–884.

Rees, D.C., DeAntonio, L. and Eisenberg, D. (1989) 'Hydrophobic organization of membrane proteins', *Science*, Vol. 245, pp.510–513.

Roder, H. and Colon, W. (1997) 'Kinetic role of early intermediates in protein folding', *Curr. Opin. Struct. Biol*., Vol. 7, pp.15–28.

Rose, G.D. and Roy, S. (1980) 'Hydrophobic basis of packing in globular proteins', *Proc. Natl. Acad. Sci. USA*, Vol. 77 pp.4643–4647.

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*, Vol. 229, pp.834–838.

Rosenbrock, H.H. (1960) 'An automatic method for finding the greatest or least value of a function', *Comput. J.*, Vol. 3, pp.175–184.

Rost, B. and Sander, C. (1993) 'Prediction of protein secondary structure at better than 70% accuracy', *J. Mol. Biol.*, Vol. 232, pp.584–599.

Rost, B., Sander C. and Schneider, R. (1994) 'Redefining the goals of protein secondary structure prediction', *J. Mol. Biol.*, Vol. 235, pp.13–26.

Roterman, I. (1995a) 'Modelling the optimal simulation path in the peptide chain folding-studies based on geometry of alanine heptapeptide', *J. Theor. Biol.*, Vol.177, pp.283–288.

Roterman, I. (1995b) 'The geometrical analysis of peptide backbone structure and its local deformations', *Biochimie*, Vol.77, pp.204–216.

Scott, M.D., van den Berg, J.J., Repka, T., Rouyer-Fessard, P., Hebbel, R.P., Beuzard, Y. and Lubin, B.H. (1993) 'Effect of excess alpha-hemoglobin chains on cellular and membrane oxidation in model beta-thalassemic erythrocytes', *J. Clin. Invest.*, Vol. 91, pp.1706–1712.

Shoichet, B.K., Baase, W.A., Kuroki, R. and Matthews, B.W. (1995) 'A relationship between protein stability and protein function', *Proc. Natl. Acad. Sci. USA*, Vol. 92, pp.452–456.

Silverman, B.D. (2001) 'Hydrophobic moments of protein structures: spatially profiling the distribution', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp.4996–5001.

Silverman, B.D. (2003) 'Hydrophobicity of transmembrane proteins: spatially profiling the distribution', *Protein Sci*., Vol. 12, pp.586–599.

Sippl, W. (2002) 'Development of biologically active compounds by combining 3D QSAR and structure-based design methods', *J. Comput. Aided Mol. Des.*, Vol. 16, pp.825–830.

Sippl, M.J., Nemethy, G. and Scheraga, H.A. (1984) 'Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O-H:O=C hydrogen bonds from packing configurations', *J. Phys. Chem*., Vol. 88, pp.6231–6233.

Skolnick, J. and Fetrow, J.S. (2000) 'From genes to protein structure and function: novel applications of computational approaches in the genomic era', *Trends Biotechnol*., Vol. 18, pp.34–39.

Song, H., Inaka, K., Maenaka, K. and Matsushima, M. (1994) 'Structural changes of active site cleft and different saccharide binding modes in human lysozyme co-crystallized with hexa-N-acetyl-chitohexaose at pH 4.0', *J. Mol. Biol.*, Vol. 244, pp.522–540.

Sosnick, T.R., Mayne, L. and Englander, S.W. (1996) 'Molecular collapse: the rate-limiting step in two-state cytochrome c folding', *Proteins*, Vol. 24, pp.413–426.

Sosnick, T.R., Shtilerman M.D., Mayne, L. and Englander, S.W. (1997) 'Ultrafast signals in protein folding and the polypeptide contracted state', *Proc. Natl. Acad. Sci. USA*, Vol. 94, pp.8545–8550.

Sosnick, T.R., Berry, R.S., Colubri, A. and Fernandez, A. (2002) 'Distinguishing foldable proteins from nonfolders: when and how do they differ?', *Proteins*, Vol. 49, pp.15–23.

Tsodikov, O.V., Record Jr., M.T. and Sergeev, Y.V. (2002) 'Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature', *J. Comput. Chem*., Vol. 23, pp.600–609.

Vojtechovsky, J., Chu, K., Berendzen, J., Sweet, R.M. and Schlichting, I. (1999) 'Crystal structures of myoglobin-ligand complexes at near-atomic resolution', *Biophys. J*., Vol. 77, pp.2153–2174.

Wallace, A.C., Borkakoti, N. and Thornton, J.M. (1997) 'TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites', *Protein Sci.*, Vol. 6, pp.2308–2323.

Wang, Z., Canagarajah, B.J., Boehm, J.C., Kassisa, S., Cobb, M.H., Young, P.R., Abdel-Meguid, S., Adams, J.L. and Goldsmith, E.J. (1998) 'Structural basis of inhibitor selectivity in MAP kinases', *Structure*, Vol. 6, pp.1117–1128.

Wang, Z., Feng, H., Landry, S.J., Maxwell, J. and Gierasch, L.M. (1999) 'Basis of substrate binding by the chaperonin GroEL', *Biochemistry*, Vol. 38, pp.12537–12546.

Warren, M.S., Brown, K.A., Farnum, M.F., Howell, E.E. and Kraut, J. (1991) 'Investigation of the functional role of tryptophan-22 in Escherichia coli dihydrofolate reductase by site-directed mutagenesis', *Biochemistry*, Vol. 30, pp.11092–11103.

Weaver, L.H., Grutter, M.G. and Matthews, B.W. (1995) 'The refined structures of goose lysozyme and its complex with a bound trisaccharide show that the 'goose-type' lysozymes lack a catalytic aspartate residue', *J. Mol. Biol*., Vol. 245, pp.54–68.

Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) 'Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores', *J. Mol. Biol*., Vol. 297, pp.233–249.

Wimley, W.C. and White, S.H. (1996) 'Experimentally determined hydrophobicity scale for proteins at membrane interfaces', *Nat. Struct. Biol*., Vol. 3, pp.842–848.

Wolfender, R., Anderson, L., Cullis, P.M. and Soulhgate, C.C. (1981) 'Affinities of amino acid side chains for solvent water', *Biochemistry*, Vol. 20, No. 4, pp.849–855.

Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. (1999) 'A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment', *Proteins*, Vol. 34, pp.220–223.

Zhang, J., Liu, Z.P., Jones, T.A., Gierasch, L.M. and Sambrook, J.F. (1992) 'Mutating the charged residues in the binding pocket of cellular retinoic acid-binding protein simultaneously reduces its binding affinity to retinoic acid and increases its thermostability', *Proteins*, Vol. 13, pp.87–99.

Zvelebil, M.J. and Sternberg, M.J. (1988) 'Analysis and prediction of the location of catalytic residues in enzymes', *Protein Eng*., Vol. 2, pp.127–138.