In Silico Biology 6 (2006) 589–600
IOS Press

Sequence-Structure-Function Relation Characterized *in silico*

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

[a]*Department of Bioinformatics and Telemedicine, Collegium Medicum – Jagiellonian University, Kopernika 17, 31-501 Krakow, Poland*
[b]*Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Krakow, Poland*
[c]*Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, Reymonta 4, 30-059 Krakow, Poland*
[d]*Chair of Biochemistry, Collegium Medicum – Jagiellonian University, Kopernika 7, 31-034 Krakow, Poland*

Edited by E. Wingender; received 27 April 2006; revised 1 August 2006; accepted 9 October 2006; published 7 November 2006

ABSTRACT: Methods for biological function recognition *in silico* appeared to be useful also for identifying characteristics of structure-to-function relations. The introduction of a three-dimensional Gauss function was assumed to represent the hydrophobic core in a protein molecule. The discrepancy between idealized "fuzzy oil-drop" and the observed one in real proteins appeared to be localized in the ligation site or in the area of biological function related part of protein molecule. The examples of proteins presented in this paper reveal that the structure-function relation can be evaluated and characterized also using the profile of the difference in value between idealized and real hydrophobicity distribution along the polypeptide chain. The specificity of particular polypeptide chain fragments in respect to their biological function and their specific participation in active site creation is discussed in this paper. The scale allowing comparison of different proteins in respect to their ligand-binding sites characteristics is introduced.

KEYWORDS: Hydrophobicity, biological function, active site, proteins of unknown biological function

INTRODUCTION

The original idea of the oil-drop model introduced by Kauzmann [Kauzmann, 1959] was applied to the model of "fuzzy oil-drop" presented in this paper. According to the oil-drop model, the highest hydrophobicity is expected to be localized in a central part of the protein molecule. The surface of the protein is expected to be covered by hydrophilic residues. The same conditions are satisfied by a three-dimensional Gauss function with its maximum in the center of the ellipsoid (origin of coordinate system) and distribution of decreasing hydrophobicity in the distance dependent form characteristic for this function treating the protein molecule as fuzzy oil-drop.

The model of discrete hydrophobicity distribution was also represented formerly by a method that partitions conformations into inner, middle and outer ellipsoidal spaces for the purpose of recognizing single well-formed hydrophobic cores. The inner core score measures the extent to which hydrophobic

*Corresponding author. E-mail: myroterm@cyf-kr.edu.pl.

Electronic publication can be found in *In Silico Biol.* **6**, 0052 <http://www.bioinfo.de/isb/2006/06/0052/>, 7 November 2006.

side-chains partition to the central region of the molecule, to the exclusion of polar and charged side-chains. The same procedure applied to the two other layers orients the side chains according to their decreasing hydrophobicity. This model was applied in [Bonneau *et al.*, 2001]. The model introduced in this paper treats the hydrophobicity distribution in continuous form.

The sole superficial location of strongly hydrophobic clusters was also examined for protein-protein interactions and used to indicate ligand binding sites [Young *et al.*, 1994].

The packing heterogeneity versus the cavity formation was examined in [Kurochkina *et al.*, 1998] to identify and classify the biological function of proteins.

The model presented in this paper not only identifies the localization of ligand binding site [Brylinski *et al.*, 2006a]. It allows also for the identification of the specific feature and nature of structure-function relation particularly in the locus of ligand binding.

The model presented in this paper was also applied to the procedure oriented on protein folding in the hydrophobic environment of three-dimensional Gauss function: TA0354_69_121 from *Thermoplasma acidophilum* target T0215 in CASP6 competition (www.predictioncenter.org) [Konieczny *et al.*, 2006], BPTI (PDB ID: 4PTI) [Brylinski *et al.*, 2006b], ribonuclease A (PDB ID: 5RAT) [Brylinski *et al.*, 2006c], lysozyme (PDB ID: 2EQL) [Brylinski *et al.*, 2006d]. The active site recognition based on the presented model was applied to following proteins: cAMP-dependent protein kinase (PDB ID: 1CDK), cyclin-dependent protein kinase 2 (PDB ID: 1E1V), proto-oncogene tyrosine-protein kinase ABL (PDB ID: 1IEP), S-lectin (PDB ID: 1SLT) [Brylinski *et al.*, 2006a].

MATERIAL and METHODS

Data

The following protein molecules were selected for the analysis presented in this paper: ferric hydroxamate uptake receptor from *E. coli* (PDB ID: 2FCP), anion-selective porin from *Comamonas acidovorans* (PDB ID: 1E45), topoisomerase IV subunit B from *E. coli* (PDB ID: 1S16), and hypothetical protein PH1257 from *P. horikoshii OT3* (PDB ID: 2D13). The ferric hydroxamate uptake receptor (a member of the class of membrane and cell surface proteins and peptides) was chosen in order to verify the fuzzy oil-drop model for "inside-out" proteins. Moreover it is noteworthy that hypothetical protein PH1257 was recently deposited in Protein Data Bank with the annotation of unknown biological function. The results obtained for this protein may be particularly useful for further experimental studies oriented on biological function identification (e.g. site-directed mutagenesis). The active site of the two proteins of known function was indicated by the localization of phosphoaminophosphonic acid adenylate ester and a magnesium ion for topoisomerase IV and ferric hydroxamate uptake receptor, respectively, according to the crystal structures.

Hydrophobic core present in real proteins

The protein under consideration localized with its geometric center in the origin of the coordinate system is assumed to represent an "observed oil-drop" with the empirical hydrophobicity distribution expressed by localization of effective atoms (an amino acid side chain may be characterized in terms of a hydrophobicity parameter). Before the empirical distribution of hydrophobicity can be calculated a protein molecule shall be oriented in a coordinate system as follows:

1. The longest distance between two effective atoms (the side chains represented by the geometrical center of all the non-hydrogen atoms present there) determines the Z-axis.
2. The Y-axis is oriented according to the longest distance between the projections of the effective atoms on the *XY* plane.

The empirical function presented by Levitt [Levitt, 1958] can be used to express the hydrophobicity density in a selected point in space (effective atom localization in particular), which collects all hydrophobic interaction according to the function:

$$\tilde{H}o_j = \begin{cases} \frac{1}{\tilde{H}o_{\mathrm{sum}}} \sum\limits_{i \in \{\mathrm{effatoms}\}} \tilde{H}_i^r \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right], & r_{ij} \leqslant c \\ 0, & r_{ij} > c \end{cases}$$

where $\tilde{H}o_j$ represents the empirical hydrophobicity value characteristic for the j-th grid point, $\tilde{H}_i^r$ represents the hydrophobicity characteristic of the i-th amino acid, r_{ij} is the distance between the j-th grid point and i-th effective atom in the amino acid, and c expresses the cutoff distance, which has a fixed value of 9.0 Å following the original paper of Levitt, 1958. $\tilde{H}o_{\mathrm{sum}}$ represents the sum of all the grid points hydrophobicity. Applying this function requires attribution of hydrophobicity parameter values to each amino acid. Many scales for residue hydrophobicity are available. Some of them are based on analysis of known protein 3D structures, while others are derived from the physicochemical properties of amino acid side chains. The selection of an appropriate scale seems crucial, so a new statistics-based hydrophobicity scale for amino acids has been created and presented in [Brylinski *et al.*, 2006c]. The hydrophobicity scale for amino acids is calculated according to the model of "fuzzy oil-drop". The values of Gauss function depending on the localization of amino acid under consideration are taken as normalized hydrophobicity parameter for particular amino acid. It means, that amino acids localized closer to the molecular center have a higher hydrophobicity than those localized on the surface. The distance dependency in this scale satisfies the Gauss function characteristics. The scale appeared highly comparable to many scales used commonly to describe hydrophobicity of amino acids [e.g., Hopp and Woods, 1981; Eisenberg *et al.*, 1982; Kyte and Doolittle, 1982; Engelman *et al.*, 1986].

Idealized "fuzzy oil-drop" form

The model of "fuzzy oil-drop" has been presented elsewhere [Konieczny *et al.*, 2006] according to which the idealized hydrophobicity distribution is represented by the three-dimensional Gauss function:

$$\tilde{H}e_j = \frac{1}{\tilde{H}e_{\mathrm{sum}}} \exp\left(\frac{-(x_i - \bar{x})^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_i - \bar{y})^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_i - \bar{z})^2}{2\sigma_z^2}\right).$$

The value of the probability distribution (as the value of the Gauss function is usually interpreted) $\tilde{H}e_j$ is assumed to represent the hydrophobicity distribution in a selected point belonging to the protein body (localization of effective atoms in particular). The hydrophobicity maximum is localized in the center of the ellipsoid and decreases in a distance-dependent manner according to the three-dimensional Gauss function. The mean value at which the Gauss function reaches its maximum is localized at the (0, 0, 0) point in a coordinate system. The standard deviation represents the size of the drop, the values of three standard deviations (along each axis) determine the size of the drop: (σ_x, σ_y, σ_z). They depend on the distribution of effective atoms' localization and on the length of the polypeptide chain under consideration. The size of the oil drop expressed by (σ_x, σ_y, σ_z) is calculated according to the standard

deviation equation for effective atoms' localization for the orientation of the molecule in the coordinate system as was defined above:

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{i \in \{\text{effatoms}\}} (\bar{x} - x_i)^2},$$

where N denotes the number of effective atoms. Values of σ_y and σ_z were obtained analogically. Coordinates (x_i, y_i, z_i) of an i-th effective atom were calculated as the mean value of coordinates for all non-hydrogen side-chain atoms of a particular i-th amino acid. The Gauss function depends on the x, y and z variables which express the positions in space (in coordinate system). Based on positions of effective atoms and their distribution the (σ_x, σ_y, σ_z) are calculated. In consequence, the distribution of effective atoms determines the size of protein molecule.

Active site recognition

In the case that both forms of hydrophobicity are calculated for common points in space (for example in a grid system or in the points representing positions of the effective atoms of side chains) and both of them are standardized (sum of all grid points hydrophobicity equals 1), the comparison between idealized and observed hydrophobicity density can be calculated. The differences between the theoretical and empirical distributions $\Delta\tilde{H}_i$ express the irregularity of hydrophobic core construction:

$$\Delta\tilde{H}_i = \tilde{H}e_i - \tilde{H}o_i$$

The theoretical fuzzy oil-drop and empirical oil-drop were calculated for all proteins taken into consideration. A color scale was introduced to express the magnitude of the difference $\Delta\tilde{H}_i$ in a particular protein area, visualizing the localization of these discrepancies in the protein molecule. The one-dimensional profiles of $\Delta\tilde{H}_i$ were smoothed by averaging of the raw data using a 5-residue running window frame. As was shown in other papers of this series [Brylinski *et al.*, 2006a, Brylinski *et al.*, 2006b, Brylinski *et al.*, 2006c, Brylinski *et al.*, 2006d, Konieczny *et al.*, 2006], the high discrepancy between idealized and observed hydrophobicity density in proteins reveals the area of function-related active site. The color scale is introduced to represent the magnitude of the $\Delta\tilde{H}_i$ (color scale can be seen on the website version of the journal available: http://www.bioinfo.de/isb/). The color scale expressing the range of discrepancy can be used for three-dimensional presentation of the protein molecule. The more red color the higher discrepancy between idealized and observed oil drop appears in a particular area of the protein molecule. In consequence the red color distinguishes the area of hydrophobicity deficiency. The second form of presentation is the $\Delta\tilde{H}_i$ profile along the polypeptide chain. Discrepancies assigned to individual amino acids were smoothed with respect to their neighbors in the sequence with the use of averaging 5 amino acids long frame.

Quantitative scale to measure differences between $\Delta\tilde{H}_i$ distributions

The visual analysis of $\Delta\tilde{H}_i$ distribution reveals the differences between profiles. The single, isolated and well defined fragment of the polypeptide chain of high $\Delta\tilde{H}_i$ suggests the active site constituted by this fragment. The highly distributed residues of positive $\Delta\tilde{H}_i$ along the polypeptide chain suggest the creation of the active site as the conjunction of events of many residues necessary to meet together in

a particular area in space. The longer linear distance (number of residues) between positions of amino acids participating in the active site, the more difficult is the creation of active site.

The quantitative scale can be introduced to measure differences between active sites (understood as the ligand binding site) characteristics based on the sequence-related-localization of amino acids participating in the active site construction (understood as high hydrophobicity deficiency). The information entropy can be used to measure the degree of difficulty of ligand binding site predictability. The analysis of positive $\Delta\tilde{H}_i$ fragments can be used to characterize the active site. On the other hand, the analysis of the negative $\Delta\tilde{H}_i$ may also give insight into active site creation. Particularly the flexibility of fragments separating positive $\Delta\tilde{H}_i$ fragments may describe the difficulty in reaching particular positions of amino acids creating the ligand binding site.

The distribution of positive $\Delta\tilde{H}_i$ fragments can be characterized as follows:

1. The information entropy can be calculated for positive $\Delta\tilde{H}_i$ fragments assuming that the size of surface between the positive $\Delta\tilde{H}_i$ profile fragment the x-axis and the number of fragments of positive $\Delta\tilde{H}_i$ as well as their dissipation can express the probability of participating in ligand binding site:

$$SH^p = -\sum_{i}^{M_p} p_i \cdot \log p_i,$$

where M_p denotes the number of sequence fragments with positive $\Delta\tilde{H}_i$ values, and

$$p_i = \sum_{j=1}^{N_{ij}} \frac{\Delta\tilde{H}_j^p}{\Delta\tilde{H}_{\text{total}}^p},$$

where N_{ij} is the number of residues belonging to the i-th fragment of positive $\Delta\tilde{H}_i$ values. The upper index p denotes the fragments of positive $\Delta\tilde{H}_i$. $\Delta\tilde{H}_{\text{total}}^p$ denotes the sum of all positive values of $\Delta\tilde{H}_i$ in a whole polypeptide chain. The p_i values are normalized to 1. The SH^p value can be expressed in bits. The higher the SH^p the more difficult is the prediction of proper fragments to participate in the ligand binding site.
2. The SH^p value depends on M_p. To make possible the comparison of SH^p values describing polypeptides of different length, one can calculate the maximum $SH_{\max}^p$ for a defined number of M_p. The closer the calculated SH^p value is versus the $SH_{\max}^p$ the more complicated is the prediction of ligand binding site. $SH_{\max}^p$ value can be calculated as follows:

$$SH_{\max}^p = -\sum_{i}^{M_p} \frac{1}{M_p} \cdot \log \frac{1}{M_p}.$$

The value of $SH_{\max}^p$ expresses the most difficult predictability, when all elements (fragments of positive $\Delta\tilde{H}_i$) are equally probable to participate in ligand binding site creation. The difference between $SH_{\max}^p$ and SH^p calculated for a particular $\Delta\tilde{H}_i$ profile may characterize the specificity of dissipation of fragments participating in ligand binding site creation.
3. The SH^p value can also be expressed on the basis of number of amino acids belonging to particular fragments of positive $\Delta\tilde{H}_i$. The p_i values in the equation for SH^p shall be substituted by relative numbers of amino acids in fragments versus the total number of all amino acids belonging to positive $\Delta\tilde{H}_i$ fragments.

4. The ligand binding site according to the analysis of the $\Delta\tilde{H}_i$ profile can also be treated as the conjunction of events understood as the occupation of positions in close mutual vicinity in space. Thus the probability P for such event can be calculated as the multiplication of probabilities expressed in the equation for p_i:

   $$I = -\log_2 P \quad \text{[bit]},$$

   where

   $$P = \prod_i^M p_i,$$

   where M denotes the number of fragments participating in active site creation with the p_i: probability calculated according to their strength (expressed by the number of amino acids in a fragment or by the integral of $\Delta\tilde{H}_i$).
   To characterize the active site in a protein the analysis of the fragments separating the positive $\Delta\tilde{H}_i$ fragments is necessary. Similar analysis (to this shown above) can be performed to describe the fragments of negative $\Delta\tilde{H}_i$ understood as those which do not participate in ligand binding site construction. The SH^m calculated in analogy to SH^p for negative (index m) $\Delta\tilde{H}_i$ characterizes the separators (in respect to the positive $\Delta\tilde{H}_i$ positions). The number of fragments of negative $\Delta\tilde{H}_i$ values influences SH^m value, the number of amino acids in separating fragment is important as well as the number of energetically acceptable conformers for amino acids in negative $\Delta\tilde{H}_i$ fragments (Table 1). These numbers have been taken from [Némethy *et al*., 1966] which, when put into the formula for P, express the influence of chain flexibility in separators.

Software

Molecular images and profiles were obtained with the program Reveal, a simple tool enabling visualization and quantitative analysis of $\Delta\tilde{H}_i$ of virtually every protein stored in PDB. Reveal is freely available online via JavaWebStart technology on www.bioinformatics.cm-uj.krakow.pl/reveal/.

RESULTS

Idealized versus observed hydrophobicity distribution

The two versions of graphic presentation applied to selected protein molecules are shown in Figs 1–4. The three-dimensional presentation of molecules under consideration reveals a high concentration of red color in the pockets treated as a ligation locus or channel in trans-membrane protein molecules. The green area suggests low or negative hydrophobicity discrepancy.

High- and low-entropy ligand binding site

The profile of $\Delta\tilde{H}_i$ distribution along the polypeptide chain reveals fragments of polypeptide chain representing high discrepancy between idealized and observed hydrophobicity distribution. The $\Delta\tilde{H}_i$ maxima localizations are of particular interest. The obvious differences between $\Delta\tilde{H}_i$ profiles seem to differentiate the active site characteristics in the compared proteins. The compact group of about 100 aa

Table 1
The number of possible conformers for particular amino acids with different precision

Amino acid	Step size		
	$1^\circ \times 1^\circ$	$5^\circ \times 5^\circ$	$10^\circ \times 10^\circ$
Pro	3221	129	32
Val	6043	242	60
Ile	6043	242	60
Cys	17646	706	176
Asn	17646	706	176
Asp	17646	706	176
Glu	17646	706	176
Gln	17646	706	176
His	17646	706	176
Leu	17646	706	176
Lys	17646	706	176
Met	17646	706	176
Arg	17646	706	176
Ser	17646	706	176
Thr	17646	706	176
Trp	17646	706	176
Tyr	17646	706	176
Phe	17646	706	176
Ala	20023	801	200
Gly	66778	2671	668

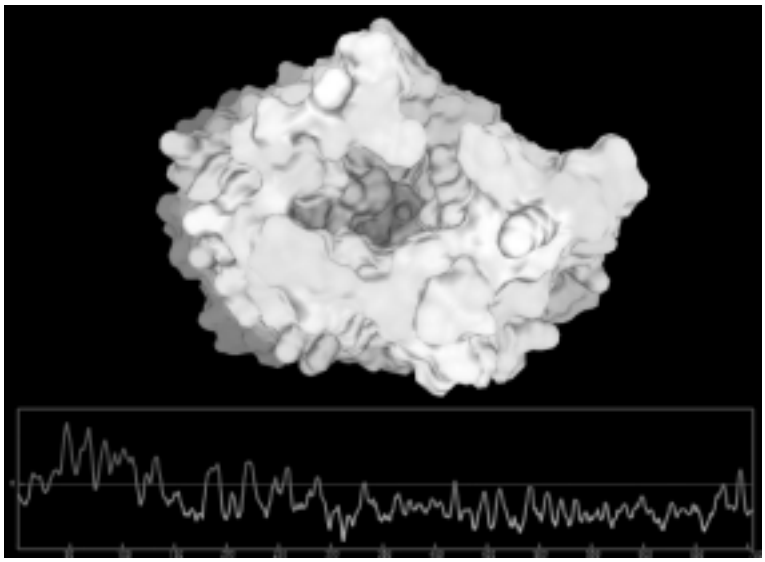

Fig. 1. The A chain of 2FCP showed and colored according to hydrophobicity discrepancy $\Delta\tilde{H}_i$. Distribution of $\Delta\tilde{H}_i$ along the sequence exhibits a compact group of amino acids with high discrepancy, revealing their location in the active-site.

in 2FCP protein seems to be responsible for active site creation, while the highly dispersed individual amino acids in protein 1S16, chain A, seem to participate in active site forming. The analysis of the $\Delta\tilde{H}_i$ profile reveals variability of high $\Delta\tilde{H}_i$ values dispersion.

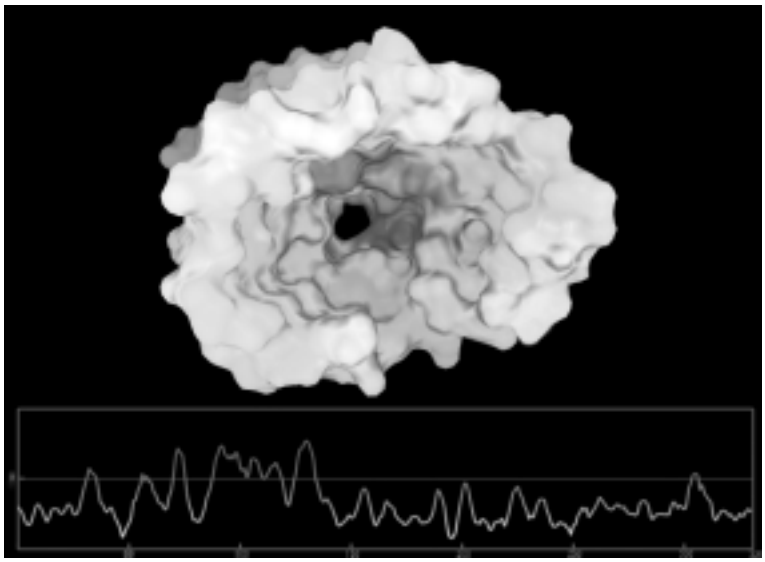

Fig. 2. Molecular surface of 1E45, the A chain, showed and colored according to hydrophobicity discrepancy. The character of $\Delta\tilde{H}_i$ distribution resembles that of 2FCP despite varying chain lengths.

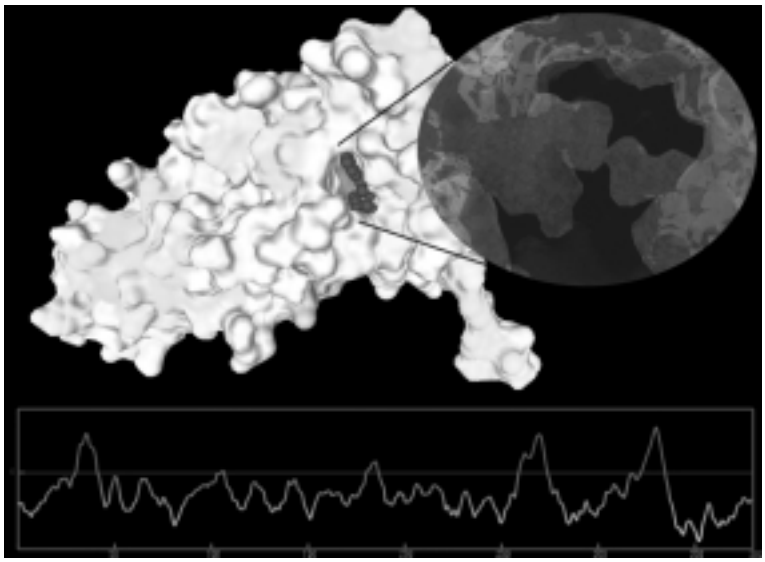

Fig. 3. Interior view of the molecular surface of 1S16. Gray-shade scale corresponds to hydrophobicity discrepancy $\Delta\tilde{H}_i$. Phosphoaminophosphonic acid-adenylate ester is visible inside the binding pocket and Mg^{2+} ion is located nearby (small dark balls). In the profile many distant regions of higher discrepancy are distinguishable, unveiling the contribution of particular amino acids to building up function-related site to be dispersed along the sequence.

The $\Delta\tilde{H}_i$ profiles can be also interpreted as different forms of participation in active site creation. The profiles shown in Figs 1 and 2 are obviously similar in a sense to the length of polypeptide participating in active site creation. The same similarity can be seen between $\Delta\tilde{H}_i$ profiles in Figs 3 and 4.

Although the selection of example proteins was arbitrary, it turned out that the proteins represented by well defined polypeptide fragments of high $\Delta\tilde{H}_i$ belong to the membrane proteins. The high $\Delta\tilde{H}_i$

Table 2
The SH^p and SH^m [bit], I^p and I^m [bit] values for proteins presented in this paper differentiating the proteins of low- and high-entropy ligand binding sites

Protein	SH^p[bit]	$SH^p_{\max}$[bit]	I^p[bit]		SH^m[bit]	$SH^p_{\max}$[bit]	I^m[bit]	
2FCP	2.28	3.58	65.9[a]	52.3[b]	2.57	3.70	64.0[c]	64.1[d]
1E54	1.83	3.00	22.2[a]	16.2[b]	1.80	3.17	47.1[c]	42.1[d]
1S16	1.79	2.32	18.0[a]	13.9[b]	2.52	2.58	15.9[c]	16.0[d]
2D13	1.83	2.58	22.2[a]	16.2[b]	2.32	2.81	24.4[c]	24.3[d]

Index p denotes positive $\Delta\tilde{H}_i$, index m denotes negative $\Delta\tilde{H}_i$ fragments.
[a] p_i calculated on the basis of integral,
[b] p_i calculated on the basis of number of residues in the fragment,
[c] p_i calculated on the basis of number of rotamers,
[d] p_i calculated on the basis of number of residues in a fragment.

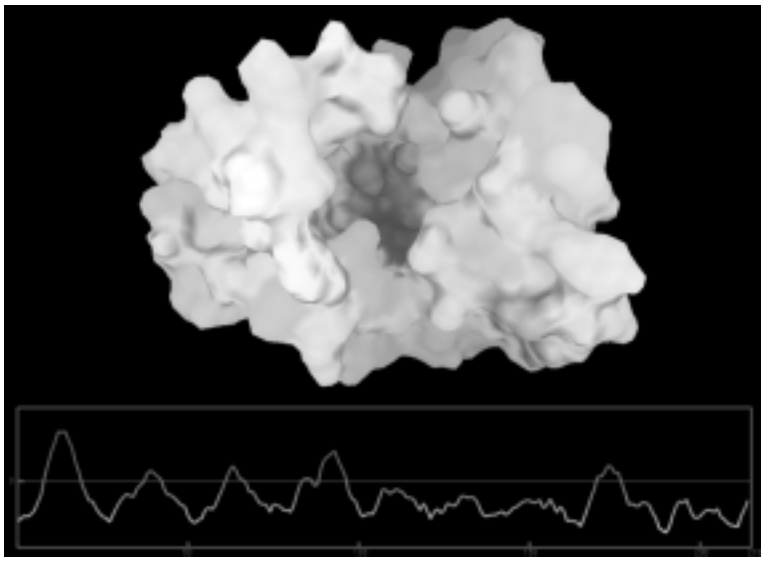

Fig. 4. Molecular surface of 2D13. According to the location of amino acids with high values of $\Delta\tilde{H}_i$ a function-related site can be expected to be placed in their close vicinity. The similarity with 1S16 in the character of the profile leads to the suggestion that both peptides might be classified in the same group when the mechanism of active-site organization is considered.

protein fragment appeared to be localized in the transport channel. Detailed analysis of the red color distribution reveals asymmetry of $\Delta\tilde{H}_i$ dispersion in the area of the transport channel of 1E54. It seems that only one site of the channel represents higher hydrophobicity than expected (oil-drop model). On the other hand, the analysis of the active site in 2FCP seems to be concentrated symmetrically along the binding pocket.

The values of SH^p and SH^m given in Table 2 allow to compare proteins of different $\Delta\tilde{H}_i$ profiles.

The results in Table 2 suggest that the high entropy ligand binding site occurs in the 1S16 protein. The SH^p and SH^m values calculated for the $\Delta\tilde{H}_i$ profile in this protein reach values close to maximum, which describe the lowest predictability of fragments participating in ligand binding site creation. The characteristics of separators between fragments of positive $\Delta\tilde{H}_i$ in this protein seems to represent the highest difficulty. The easiest to be predicted are the active sites in proteins 2FCP and 1E54. The active site in the protein 2FCP is created by short, well defined fragment of polypeptide of continuous form. The active site of protein 1S16 seems to be created by residues distributed along the whole polypeptide

chain. This is why its ligand binding site is of low predictability. The I^m values suggest that the number of rotamers is strongly related to the number of residues except the 1E54 protein. The sequence of this protein contains an exceptionally high percentage of glycines (16.31%) in comparison to 8–9% in other proteins presented in Table 2. 49 Glycines of a total of 54 of them are present in fragments of negative $\Delta\tilde{H}_i$. This is why the amount of information [bit] calculated on the basis of the number of rotamers present in separators is significantly higher in this protein.

DISCUSSION

Proteins representing different biological activity were compared in this paper. The comparison was based on the characteristics of discrepancy between idealized and empirically observed hydrophobic oil-drop form. The differences were recognized according to the form of $\Delta\tilde{H}_i$ profile along the polypeptide chain. The highly concentrated (low distribution) positions of $\Delta\tilde{H}$ maxima in the profiles of 2FCP and 1E54 appeared to determine the amino acids localized in the active centers of these proteins. The other two proteins are characterized by the profile with very short dispersed fragments of high $\Delta\tilde{H}_i$ values. The relation between structure and active site construction can be recognized in this way. The comparison of presented $\Delta\tilde{H}_i$ profiles seems to be a suitable criteria for active site classification. This type of tool seems to be very important nowadays for biological activity recognition of proteins of unknown biological function.

A quantitative scale measuring low- and high difficulty in ligand binding site creation was introduced. The values of SH^p and SH^m put in the ranking order are able to differentiate the proteins of high- and low-entropy active sites. The higher SH^p and SH^m value the more difficult to approach the residues to participate in the active site creation. The SH^p and SH^m values depending on the polypeptide chain length express the objective high difficulty in active site creation only when amino acids localized in different parts of the polypeptide are highly distributed along the polypeptide chain.

The interpretation of $\Delta\tilde{H}_i$ profiles may have a large impact on the mutations analysis. It seems that mutation in the second half of the polypeptide chain in the case of 2FCP is not very important in a sense of active site creation. Meanwhile the mutation in the same (relative) position in the polypeptide chain in the case of 1S16 seems to have a much more important impact on the active site construction.

The scale introduced for active site (or ligand binding site) characteristics as well as the $\Delta\tilde{H}_i$ profile can be created in a fully automated form. The long list of proteins (particularly those of unknown biological activity) can be ordered according to the score expressed by SH^p and SH^m values. The SH^p and SH^m values depend on the length of the polypeptide. The comparison of SH^p with $SH^p_{\max}$ (the lowest predictability) may additionally classify the range of difficulties in active site creation. The same analysis of SH^m and $SH^m_{\max}$ may additionally characterize the difficulty of active center creation from the point of view of separators. The closer the SH^m value to the $SH^m_{\max}$ the more unpredictable is the structure of separators between maxima determining the participation of residues in active site and the more difficult is to bring together the residues creating active site.

The hydrophobic deficiency $\tilde{H}e_i - \tilde{H}o_i$ has been analyzed and interpreted in this presentation to localize ligand binding cavity. The residues of $\Delta\tilde{H}_i$ higher then expected may localize the possible protein-protein interaction area. Such interaction is mostly interpreted as based on hydrophic interaction between two interacting proteins. This interpretation is planned to be tested in the next CAPRI competition. Particularly high negative $\Delta\tilde{H}_i$ or long polypeptide chain fragments of such characteristics may suggest the membrane protein as it is in the long C-terminal fragment of 2FCP protein of negative $\Delta\tilde{H}_i$ values. In this case the surface of protein can not fit well to the oil-drop model of expected zero hydrophobicity on

the surface. The 2D13 protein seems also be related to membrane representing highest negative values. The method presented can not be applied to multi-domain proteins or to aggregates of many proteins in complexes. Domains identification or protein-protein complexes recognition preceding the analysis presented in this paper seems to be necessary. It is expected that some individual proteins may not be recognized correctly on the basis of the presented method. Such examples may be interesting objects for other than the presented models for active site creation and their recognition.

It seems also, that this very simple analysis may be applied automatically to a massive comparative analysis of relation between structure and function particularly when applied to proteins of unknown biological function (expression of genes recognized numerically in human genome) [Skolnick *et al.*, 2000]. The characteristics of $\Delta\tilde{H}_i$ profiles, their classification, clustering into categories is planned to be used as criterion for biological function identification particularly to proteins of unknown biological function is planned to be generalized and unified. This work is in progress.

This model will be applied also to large (massive) scale calculation oriented on the "never born proteins" searching for those of them being of pharmacological interest. This is the subject of a grant of FP6 of European Commission of Science (grant EUCHINA Grid) (www.euchinagrid.org).

REFERENCES

- Bonneau, R., Strauss, C. E. M. and Baker, D. (2001). Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins **43**, 1-11.
- Brylinski, M., Konieczny, L. and Roterman, I. (2004). SPI – Structure predictability index for protein sequences. In Silico Biol. **5**, 0022.
- Brylinski, M., Konieczny, L. and Roterman, I. (2006a). Ligation site in proteins recognized in silico. Bioinformation **1**, 127-129.
- Brylinski, M., Konieczny, L. and Roterman, I. (2006b). Hydrophobic collapse in late-stage folding (in silico) of bovine pancreatic trypsin inhibitor. Biochimie **88**, 1229-1239.
- Brylinski, M., Konieczny, L. and Roterman, I. (2006c). Hydrophobic collapse in (*in silico*) protein folding. Comput. Biol. Chem. **30**, 255-267.
- Brylinski, M., Konieczny, L. and Roterman, I. (2006d). *Fuzzy-oil-drop* hydrophobic force field – a model to represent late-stage folding (*in silico*) of lysozyme. J. Biomol. Struct. Dyn. **23**, 519-527.
- Eisenberg, D., Weiss, R. M., Terwilliger, T. C. and Wilcox, W. (1982). Hydrophobic moments of protein structure. Faraday Symp. Chem. Soc. **17**, 109-120.
- 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. **179**(1), 125-142.
- 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. **15**, 321-353.
- Hopp, T. P. and Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA **78**, 3824-3828.
- Janin, J. (1979). Surface and inside volumes in globular proteins. Nature **277**, 491-492.
- Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. Adv. Protein Chem. **14**, 1-63.
- Konieczny, L., Brylinski, M. and Roterman, I. (2006). Gauss-function-based model of hydrophobicity density in proteins. In Silico Biol. **6**, 0002.
- Kurochkina, N. and Privalov, G. (1998). Heterogeneity of packing: structural approach. Protein Sci. **7**, 897-905.
- Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**(1) 105-132.
- Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. **104**, 59-107.
- Némethy, G., Leach, S. J. and Scheraga, H. A. (1966). The influence of amino acid side chains on the free energy of helix-coil transitions. J. Phys. Chem. **70**, 998-1004.
- Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. Science **229**, 834-838.
- Skolnick, J., Fetrow, J. S. and Kolinski, A. (2000). Structural genomics and its importace for gene functional analysis. Nat. Biotech. **18**, 283-287.

• Young, L., Jernigan, R. L. and Covell, D. G. (1994). A role of surface hydrophobicity in protein-protein recognition. Protein Science **3**, 717-729.