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

Structure-Function Relation in Proteins, 2009: 105-127 ISBN: 978-81-7895-409-7
Editor: Irena Roterman

5 Active site recognition *in silico*

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

Abstract

The biological activity of protein molecule is critical for the structural analysis. Mutations may cause the disappearing of biological function. The misfolding of polypeptide chain eliminates also the specific function characteristic for particular protein. Thus the recognition of the protein biological activity seems to be of significant importance for protein identification. Especially nowadays, when more and more proteins appear in PDB as products of structural genomics. The proteins the genes of which have been recognized using the bioinformation tools (GenScan) get synthesized in bacteria and crystallized although their biological function remains unknown. If the

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

assumption that the hydrophobicity deficiency cavity may represent the biological function-related-area is correct, the residues representing the $\Delta\tilde{H}_j$ *maxima (the value measuring the hydrophobicity irregularity calculated versus the idealized "fuzzy-oil-drop" distribution) may be interpreted as potential residues responsible for biological function, particularly when localized in close mutual vicinity. The applicability of "fuzzy-oil-drop" model for the biological activity recognition understood as ligand binding cavity identification is presented in this chapter. The "fuzzy-oil-drop" model appeared to be also useful for the identification of structural (and functional) consequences of mutations and as the tool for structural/functional similarity search in proteins. The proteins, which seem to be folded according the "fuzzy-oil-drop" model were found in the group of antifreeze proteins.*

Introduction

The comparison of the hydrophobicity distribution as it appears in the crystal form of particular protein with the idealized one (according to 3-D Gauss function) reveals the characteristics which seem to be quite important for the "*fuzzy-oil-drop*" model applicability. According to the "*fuzzy-oil-drop*" model the hydrophobic interaction assumed to stabilize the tertiary structure of protein was expected to be distributed according to the three-dimensional Gauss function (the highest concentration of hydrophobicity in the central part of the molecule with negligibly small or none hydrophobicity on the protein surface). This interpretation of hydrophobicity distribution in protein molecule is the modification of traditional "oil-drop" model introduced by Kauzman [1]. The comparison of idealized hydrophobicity (calculated according to three-dimensional Gauss function) distribution with the observed one (calculated according to Levitt's function [2] and measured by $\Delta\tilde{H}_j$ (difference between idealized and observed distribution of hydrophobicity) in real proteins revealed some irregularities. The area of high irregularity expressed as hydrophobicity deficiency (as well as in form of hydrophobicity excess) appeared to be localized in one common well defined area of protein molecule. The area of hydropbobicity deficiency was recognized to be very often the ligand binding site or active site (in enzymes).

In conclusion one may ask to what extent the identification of residues of $\Delta\tilde{H}_j$ maxima (when localized in a common area) may be treated as tool for active site recognition in proteins.

The search for the answer to this question is the main subject of this chapter.

Ligand binding site recognition

The set of proteins crystalized in form of complex with the specific ligand related to biological function or with inhibitor (enzymes) were selected from PDB to verify the hypothesis that the hydrophobicity deficiency area (versus the idealized hydrophobicity distribution) may be used to identify the biological activity of the protein. The list of proteins taken as examples are given in Tab.1.

Table 1. The list of proteins selected to verify the applicability of the procedure oriented on active site (ligand binding site) identification. The given biological activity is defined according to SCOP (Strutural Classification Of Proteins [16]).

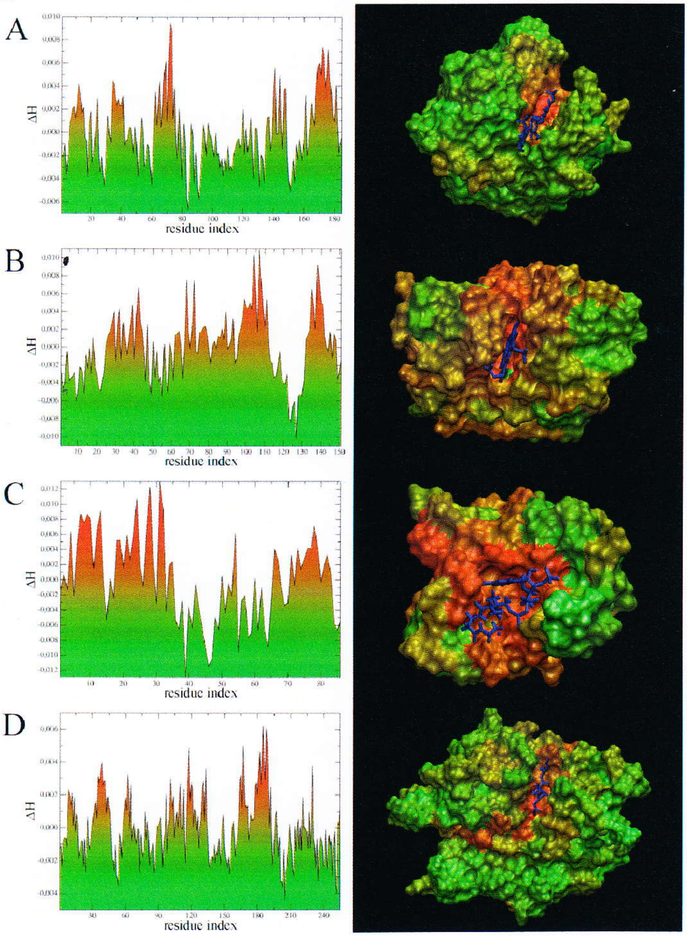
Protein	Organizm	PDB ID	Biological function
Lysozyme	*Anser anser*	154L [3]	Catabolism of bacterial membranes
Mioglobin	*Physeter catodon*	1A6M [4]	Oxygen transport
Protein binding Acetyl-CoA	*Bos Taurus*	1ACA [5]	Acetyl-CoA
TNF-α Convertase	*Homo sapiens*	1BKC [6]	Metalopeptidase
Lysozyme	*Homo sapiens*	1LZR [7]	Catabolism
Ribonuclease A	*Bos Taurus*	1RND [8]	Nuclease
Dihydropholate reductase	*E. coli*	3DRC [9]	Reductase
Mitogen-activated protein kinase	*Homo sapiens*	1A9U [10]	Phosphorylation
CDK6 kinase	*Homo sapiens*	1BLX [11]	Phosphorylation
cAMP-dependent protein kinase	*Sus Scrofa*	1CDK [12]	Kinase
cyclin-dependent protein kinase	*Homo sapiens*	1E1V [13]	6-O-cyclohexylmethyl guanine
proto-oncogene tyrosine-protein kinase ABL	*Mus musculus*	1IEP [14]	kinase
S-lectin	*Synthetic product*	1SLT [15]	lectin

The differences between theoretical (idealized) hydrophobicity density distribution and the one observed in real protein (depending on the spatial localization of hydrophobic/hydrophilic) residues in the protein body calculated as follows:

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

express the magnitude of irregularity. The $\tilde{H}t_i$ represents the theoretical hydrophobicity density as calculated according to the idealized distribution

Figure 1



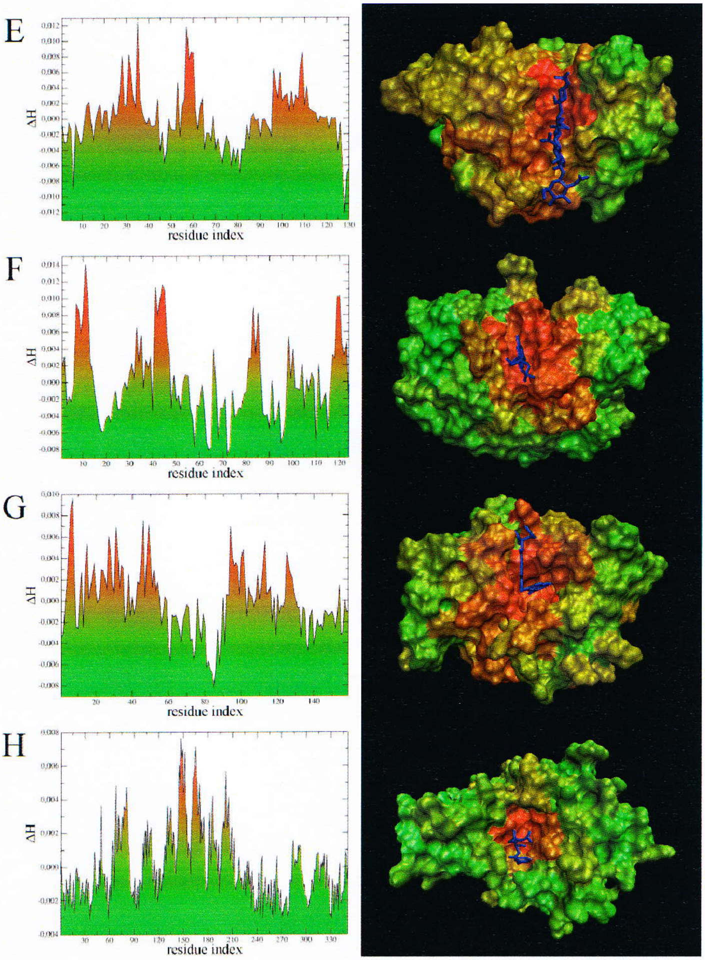


Figure 1. The $\Delta\tilde{H}_i$ profiles and 3-D presentation of hydrophobicity deficiency/excess distribution in following proteins: A – lysozyme (bacteria), B – mioglobine, C-acetyl-CoA binding protein, D – TNF-α convertase, E – lysozyme (human), F – ribonuclease A (bovine), G – protein kinase (human), H – CDK6 kinase (human). The color scale shown on $\Delta\tilde{H}_i$ profile is applied in 3-D presentation to show the localization of the residues representing the local $\Delta\tilde{H}_i$ maxima. The ligands molecules shown in dark blue color. (reproduction with permission of *International Journal of Bioinformatics research and Applications* Editor in Chief [17]).

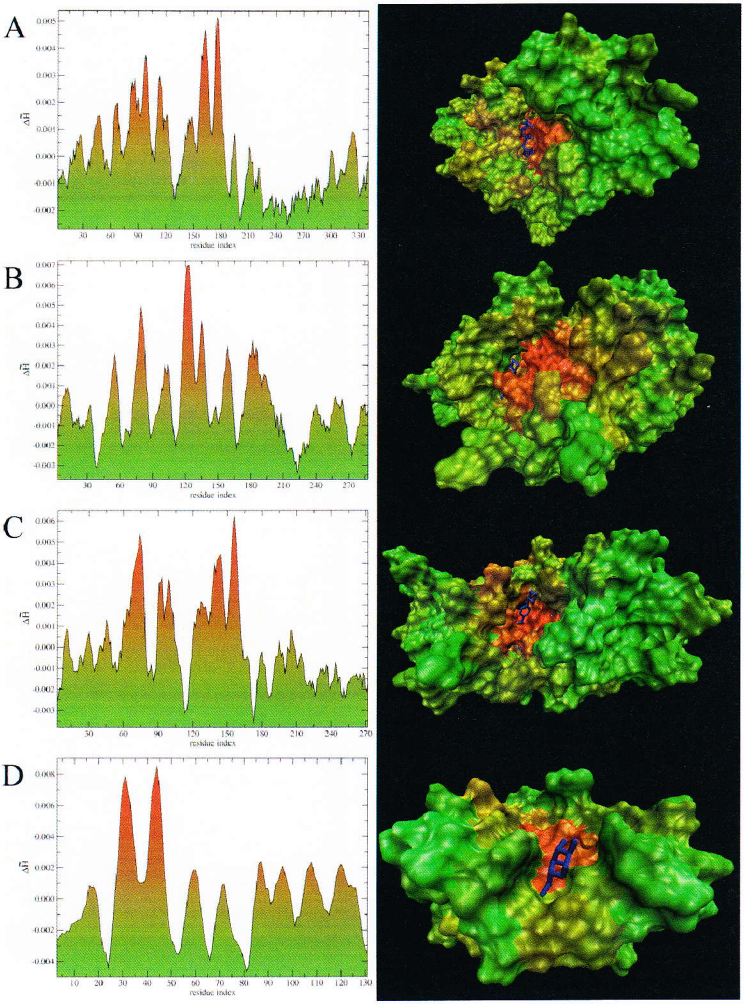


Figure 2. The $\Delta\tilde{H}_i$ profiles of proteins and spatial distribution of hydrophobicity irregularity for the ligand-protein complexes: A – cAMP-dependent protein kinase complexed with 5'adenyly-imido-triphosphate, B – cyclin-dependent protein kinase 2 complexed with 6-O-cyclohexylmethyl guanine, C – proto-oncogene tyrosine-protein kinase ABL complexed with STI-571, D – S-lectin complexed with D-galactose. The ligands are presented in dark blue color (reproduction with permission of *Bioinformation* Editor in Chief [18]).

described by 3-D Gauss function and $\tilde{Ho}_i$ represents the observed hydrophobicity density as calculated on the basis of the localization of hydrophobic/hydrophilic residues in the protein body. The $\tilde{Ho}_i$ collects the hydrophobic interaction in the distance below 9Å (cutoff distance for hydrophobic interaction according to [2]). The quantity $\Delta\tilde{H}_i$ when calculated for each residue characterizes its relative accordance to the idealized form. The high $\Delta\tilde{H}_i$ value represents the residue of hydrophobicity lower than expected. The large negative $\Delta\tilde{H}_i$ value represents the residues of hydrophobicity higher than expected. When localized on the surface of molecule may probably represent the potential area for protein-protein interaction.

The residues representing the local maxima on the profiles shown in Fig.1.and Fig.2. appeared to be localized in a mutual close vicinity and additionally in the locus of cavity. This cavity appeared to be very often occupied by ligand molecule (or inhibitor for enzymes).

The visual analysis of the relation between ligand position and high hydrophobicity deficiency localization shown in Fig.1.and Fig.2. suggests the "*fuzzy-oil-drop*" model to be the possible tool for biological function recognition. The ligand molecule has been found to occupy the cavity characterized by the hydrophobicity deficiency. It means, that the residues representing local $\Delta\tilde{H}_i$ maxima on the $\Delta\tilde{H}_i$ profile may suggest the ligand binding cavity. $\Delta\tilde{H}_i$

Quantitative measurements of active site

Assuming the residues representing local $\Delta\tilde{H}_j$ maxima which meet together in space (close mutual vicinity) create the active site, the quantitative measurements estimating the difficulty of such active site generation can be introduced. The calculation presented below may help to estimate the predictability of active site on the basis of elements of information theory.

The amount of information *I* according to Shannon definition depends on the probability *p* of the event under consideration [19].

$$I_i = -\log_2 p_i \text{ [bit]} \qquad \text{[eq.1.]}$$

Assuming that the magnitude of the $\Delta\tilde{H}_j$ is proportional to the probability of participating in active site generation the amount of information carried by one fragment of positive $\Delta\tilde{H}_j$ fragment can be calculated as follows:

$$I_j = -\log_2 \sum_{i=1}^{k} p_i \qquad \text{[eq.2.]}$$

The *j-th* fragment of positive $\Delta\tilde{H}_j$ composed of *k* amino acids is carrying the amount of information expressed by the eq.1. This assumption is correct on condition that all positive $\Delta\tilde{H}_j$ values are standardized as follows:

$$\sum_{i=1}^{K} \sum_{j=1}^{ij} p_{i,j} = 1.0$$

Where K - number of fragments of positive $\Delta\tilde{H}_j$ of $h_{i,j}$ each residue, J_i – number of residues belonging to *i-th* fragment.

The active site requires the meeting of all positive $\Delta\tilde{H}_j$ fragments which is the event of the character of conjunction. Thus the probability to put together all fragments of positive $\Delta\tilde{H}_j$ May be calculated as follows:

$$P = \prod_{i=1}^{K} p_i \qquad \text{[eq.3.]}$$

Where *P* denotes the probability of all positive fragments *(i=1 to K)* to meet together. In consequence the amount of information necessary to generate the particular active site can be calculated as follows:

$$I = -\log_2 P \qquad \text{[eq.4.]}$$

All quantities expressing amount of information are given in bits.

Another quantity which may be also calculated to characterize the active site is the information entropy. Its value may measure the level of complexity of the active site generation.

The *SE* (information entropy also expressed in bits) can be calculated as follows according to Shannon definition [19]):

$$SE = -\sum_{i}^{K} P_i \log_2 P_i [bit] \qquad \text{[eq.5.]}$$

Where *SE* – information (which may be interpreted as structural) entropy, P_i probability of *i-th* fragment to be a part of the active site (equal to the sum of probabilities for all residues participating in the *i-th* fragment of positive $\Delta\tilde{H}_j$).

The value of *SE* may be interpreted as follows:

Assuming that only one fragment of positive $\Delta\tilde{H}_j$ is present in a whole polypeptide chain. The uncertainty of the prediction that these residues may meet together is equal to 1. It means that the prediction is certain. The event of p=1. does not carry any information. This is the deterministic case (Fig.3.A). Other words – it is very easy (obvious) to predict the construction of the active site.

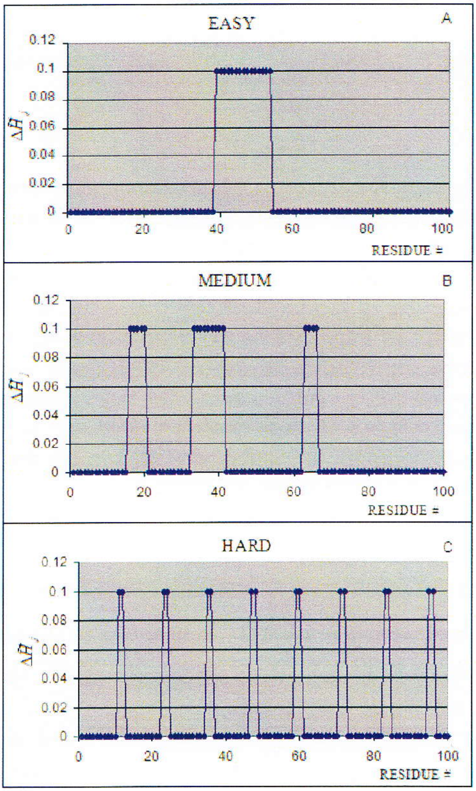


Figure 3. The examples visualizing the easy (A), medium (B) and hard (C) solutions. The example called easy represents the deterministic case due to the only one possibility of generating the active site created by the fragment of high $\Delta\tilde{H}_j$. The hardest situation happens when many possibilities of equal probability (expressed by $\Delta\tilde{H}_j$ values) represent particular solutions for elementary event.

The opposite case may be presented as follows: many short fragments of positive $\Delta\tilde{H}_j$ participate in active site generation (Fig.3.C). Predictability is difficult in such a case. *SE* parameter takes large value. *SE* value depends on the number of elements (fragments) participating in the active site generation. *SE* depends also on the length of polypeptide.

SE value is able to express the degree of difficulty of active site generation.

The situation shown as B in Fig.3. is the most frequent, when few fragments of different probability ($\Delta\tilde{H}_j$) participate in active site generation.

The dependency on the length and/or number of fragments may be eliminated in comparison of the polypeptides of different length calculating the relative *SE*.

Its value :

$$SE_{REL} = SE/SE_{MAX}$$

where SE_{MAX} expresses the situation when *K* fragments are equally probable to participate in active site generation (Fig.3.C). This is the representation of random solution, which is characterized by the highest *SE* value for particular number of fragments (and particular polypeptide chain length). The random distribution of elements is the most difficult case to take the decision – (in our example - to meet all fragments in the close mutual vicinity).

SE_{REL} expresses the relative "distance" of the case under consideration versus the maximum SE_{MAX} (representing the fully random situation). The larger is the distance the less random character is present in particular case.

It is possible to compare the active sites calculating the *SE* values and/or SE_{REL}.

This is why the proteins presented in this and other chapters are characterized using *SE* parameters.

The fragments of positive $\Delta\tilde{H}_j$ are responsible for active site generation, while the fragments of negative $\Delta\tilde{H}_j$ values are expected to generate the areas responsible for protein-protein complexation (particularly when occurring on the surface of the protein).

This is why the *SE* parameters in form SE_+ and SE_- describing the fragments of positive and negative $\Delta\tilde{H}_j$ values is calculated.

Protein-protein complexes identification

The ligand binding site is recognized as hydrophobicity deficiency cavity expressed by the local $\Delta\tilde{H}_j$ maxima. The complexation of two proteins may follow the same mechanism as applied for protein-ligand complex generation.

Such case may be observed in crystal of the P19INK4D-kinaze with the inhibitor CDK6 [12]. The complexation of these two proteins is shown in Fig.4.

The $\Delta\tilde{H}_j$ profile for P19INK4D inhibitor (Fig.4.A.) and its 3-D presentation with the P19INK4D inhibitorCDK6-kinase treated as ligand (presentation in dark blue). The presentation vice-versa is given on two left pictures (Fig.4.B.) where the inhibitor plays the role of ligand and kinaze is the target molecule. These two proteins generate the complex playing mutually the role of ligands occupying the binding cavity (hydrophobicity deficiency area). The color scale applied according to the color scale in Fig.1 and 2.

The recognition of ligand binding site on the basis of "*fuzzy oil drop*" model has been also presented in [20] for other proteins (1A6M, subtilisin - 1BH6, carboxypeptidase A2 - 1DTD, chymotrypsin – 1GG6, c-type lysozyme – 1LMQ, and ribonuclease 1RGE) and applied for protein-protein recognition (transcriptional antiterminator LicT – 1H99, cohesion-dockerin complex – 1OHZ, serine/threonine phosphatase-1 – 1S70).

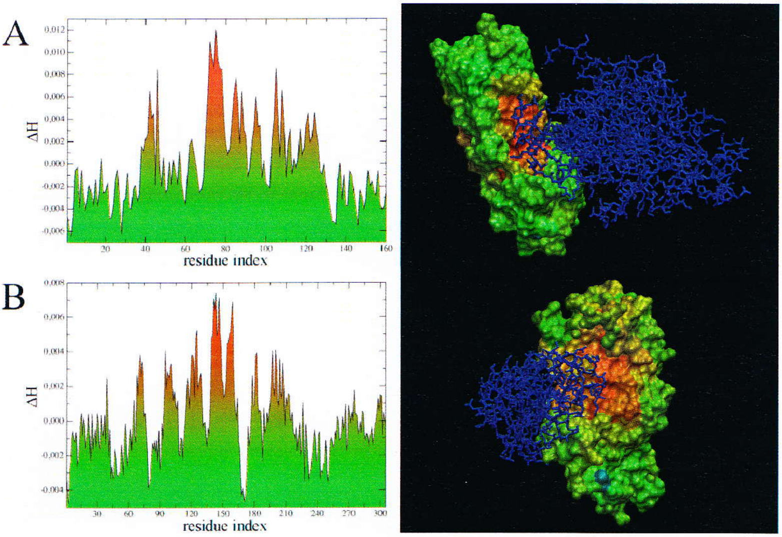


Figure 4. The kinase complex with inhibitor. The two proteins play mutually the role of target and ligand. The $\Delta\tilde{H}_j$ profile for CDK6 – kinase (upper left profile) with 3-D presentation of this protein (space filling model) (lower left) with the P19INK4D inhibitor treated as ligand (presentation in dark blue) (reproduction with permission of *International Journal of Bioinformatics research and Applications* Editor in Chief [17])

Recognition of biological function of proteins of "unknown function" status

Assuming that the residues of $\Delta\tilde{H}_j$ local maxima generate the active site, the biological function of proteins of "unknown function" status (according to PDB classification) may be recognized. The probable localization of potential active site (or ligand binding site) may be identified using $\Delta\tilde{H}_j$ profiles. The list of proteins taken as examples is given in Tab.2. The $\Delta\tilde{H}_j$ profiles of proteins listed in Tab.2. are presented in Fig.5.

Table. 2. The selected proteins deposited in PDB with the Unknown Function status.

PDB ID	Research Group
2CV9 [21]	RIKEN Structural Genomics/Proteomics Initiative (RSGI)
2FBL [22]	Midwest Center for Structural Genomics (MCSG)
2FFI [23]	Northeast Structural Genomics Consortium (NESG)

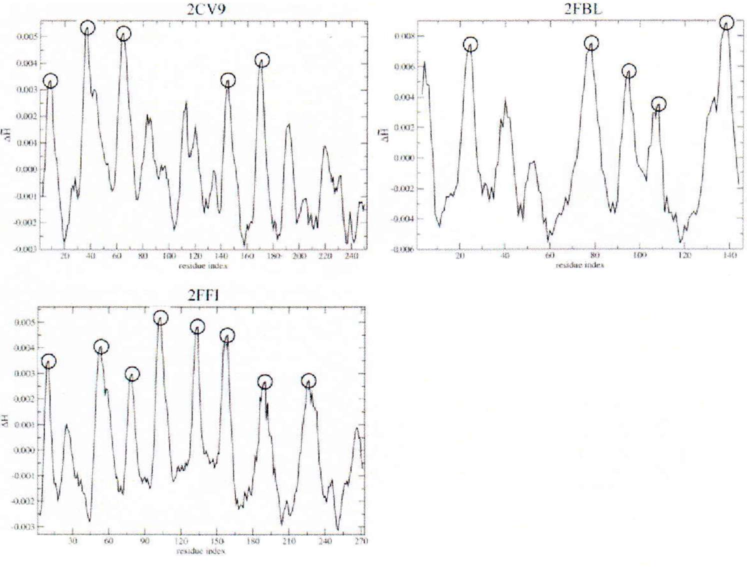


Figure 5. The $\Delta\tilde{H}_j$ profiles for three proteins of unknown biological structure : 2CV9 (Riken Structural Genomics/Proteomics Initiative (*RSGI*)), 2FBL (Midwest Center for Structural Genomics (*MCSG*)) and 2FFI (Northeast Structural Genomics Consortium (*NESG*)).

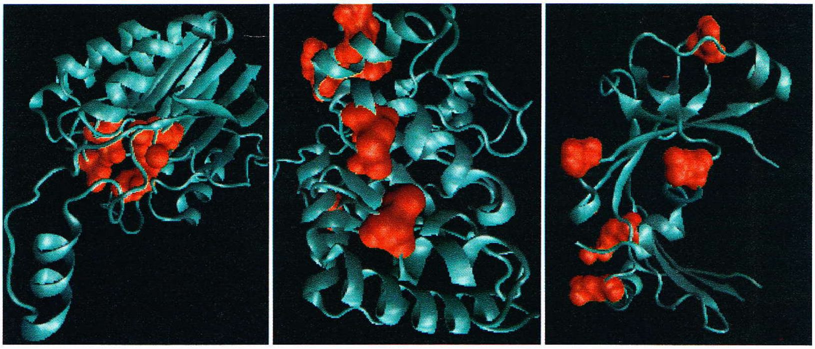


Figure 6. The 3-D presentation showing the localization of the residues of local $\Delta\tilde{H}_j$ maxima. The proteins 2CV9 and 2FFI (two pictures on left) represent quite good concentration of residues of high $\Delta\tilde{H}_j$ suggesting the active site in the distinguished area although the residues of high $\Delta\tilde{H}_j$ in 2FBL (the picture on right) seem to be quite distributed making the identification of active site difficult.

The analysis of the pictures shown in Fig.6. suggests that the quite tight packed residues representing $\Delta\tilde{H}_j$ maxima may probably represent the active site (or ligand binding site) in 2CV9, although the larger distribution of such residues in 2FBL may rise some doubts.

The extended analysis taking many proteins under consideration is necessary to define the conditions and limits for "*fuzzy-oil-drop*" model applicability as the tool for active site recognition *in silico*.

Similarity search

The $\Delta\tilde{H}_j$ profiles appear to be quite differentiated and strongly dependent on the protein specificity (Fig.7.). It seems possible to take the comparison of $\Delta\tilde{H}_j$ profiles as the tool to search for proteins similarity. It may express the structural as well as functional similarity assuming that the dispersion of hydrophobicity deficiency and/or excess is specific and function related.

The large set of proteins (above 400) was analyzed to evaluate the applicability of "*fuzzy-oil-drop*" model for different proteins. Some of them appeared to represent the identical or very similar $\Delta\tilde{H}_j$ profiles suggesting the similarity of proteins under consideration.

The selected proteins appeared to represent the similar *SE* parameters making probable the recognition of the biological function of one of them (2CRE - unknown function status) [24]. The analysis of sequence comparison pointed three candidates to be similar to the 2CRE protein (Tab.3.).

The protein 2CRE appeared similar to the 1U5S-A according to *SE* parameters (Tab.4.). Almost all parameters represent the lowest difference in pair-wise comparison.

Table 3. The sequence similarity between three proteins Three of them (2DA9 [25], 1U5S-A [26] and 1U5S-B [26]) of known biological activity compared to the one 2CRE of unknown biological function.

PROTEINS	**2CRE – 2DA9**	**2CRE – 1U5S-A**	**2CRE – 1U5S-B**
Identity	42.86 %	21.43 %	15.71 %
Similarity	51.43 %	35.71 %	18.57 %
Number of gaps	1	4	5
Score (ClustalW)	42	22	3

Table 4. The SE characteristics for proteins recognized as similar to the 2CRE protein.

PROTEIN	SE_{+}	SE_{+max}	SE_{+REL}	I_{+}	SE_{-}	SE_{-max}	SE_{-REL}	I_{-}
2CRE	3.25	4.00	0.17	57.23	3.40	4.09	0.17	54.92
2DA9	3.15	3.33	0.05	35.34	2.88	3.46	0.16	45.32
1U5S-A	3.36	4.00	0.16	61.50	3.43	4.00	0.14	54.44
1U5S-B	1.96	3.00	0.35	26.71	2.29	3.17	0.27	36.41

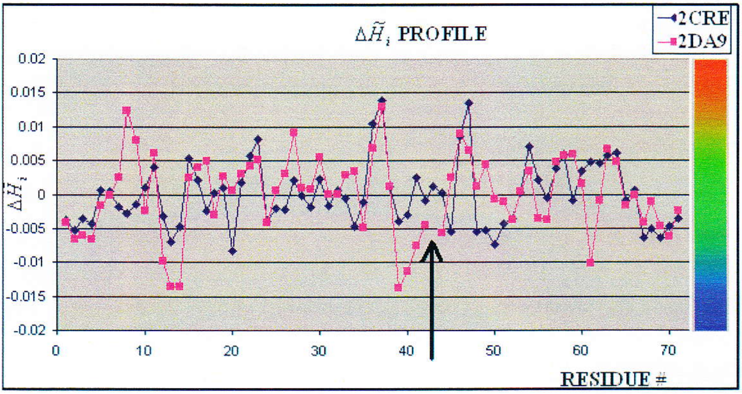


Figure 7. The similarity of $\Delta\tilde{H}_j$ profiles of 2CRE (unknown function status) and 2DA9 (known biological function). The arrow indicates the insertion locus. The color scale shows the colors applied for 3-D presentation of hydrophobicity irregularity.

The $\Delta\tilde{H}_j$ profiles for 2CRE and 2DA9 reveals that one deletion/insertion (identified also according to sequence alignment) makes these two profiles highly similar.

The $\Delta\tilde{H}_j$ profiles of 2CRE and 2DA9 visualize the similarity expressed quantitatively by the *SE* parameters as well as their 3-dimensional (Fig.8). The distribution of hydrophobicity deficiency/excess in 3-D presentation reveals quite high similarity suggesting the similar function of these two proteins. The gap, which was recognized in sequence comparison (according to ClustalW) is also seen in $\Delta\tilde{H}_j$ profiles making two profiles more similar.

Additionally it turned out that both proteins represent structural similarity of SH3 domain form (also the 1U5S-A represent the SH3 motif).

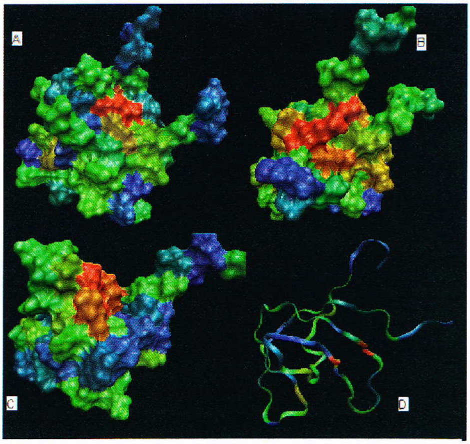


Figure 8. The 3-D presentation of hydrophobicity deficiency/excess (color scale according to Fig.7.) A – 2CRE (unknown function), B – 2DA9 – known biological function, C – 1U5S-A (known biological function), D – 2CRE – in ribbon presentation (colored according to magnitude of $\Delta\tilde{H}_j$ values) showing the typical SH3 domain form.

Proteins the structures of which are accordant with the "*fuzzy-oil-drop*" model

The next question, which can be asked in relation to "*fuzzy-oil-drop*" model applicability is whether there are proteins of the hydrophobicity distribution accordant to the idealized (three-dimensional Gauss function) one as applied in "*fuzzy-oil-drop*" model.

The large set of proteins (above 400) of 70 amino acids in a polypeptide chain has been analyzed [27].

The group of proteins representing the antifreeze proteins was present in that data base.

The biological function of these proteins is to be distributed in the organism preventing the water to change into the ice. The molecule playing this role shall be very well soluble and interact with proteins although no high specificity in protein-protein complex generation is expected.

To satisfy these expectations the hydrophobicity excess is expected to be present on the surface of the antifreeze protein to prevent the large scale ordering of water molecules in the area near the protein surface and not necessarily any ligand binding cavity is expected as function-related characteristics. No specific binding cavity is expected in the protein playing the role of antifreeze.

The molecule 1MSI [28] (north Atlantic ocean pout *Macrozoarces americanus*) and its mutant 1KDE [29] (97.14% sequence similarity (97.14% identity) represent very similar $\Delta\tilde{H}_j$ profiles (almost identical) (Fig. 9.) and the hydrophobicity distribution in 3-D representation (Fig.10.).

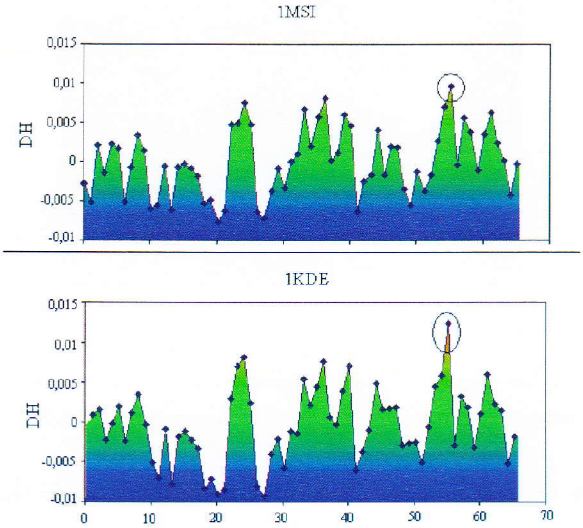


Figure 9. The $\Delta\tilde{H}_j$ profiles of antifreeze proteins 1KDE and 1MSI. The color scale visualizing the ΔH_j (DH) value is applied for three-dimensional distribution of $\Delta\tilde{H}_j$ (see Fig.10.). The only slightly red residues (number 55) represent the hydrophobicity deficiency are distinguished by circles.

The single one red (high hydrophobicity deficiency) residue present in $\Delta\tilde{H}_j$ profile is almost entirely buried in the protein body. The surface of these proteins is covered by green – (accordant with the "*fuzzy oil drop*" hydrophobicity distribution) and blue fragments - (higher than expected hydrophobicity exposed on the protein surface).

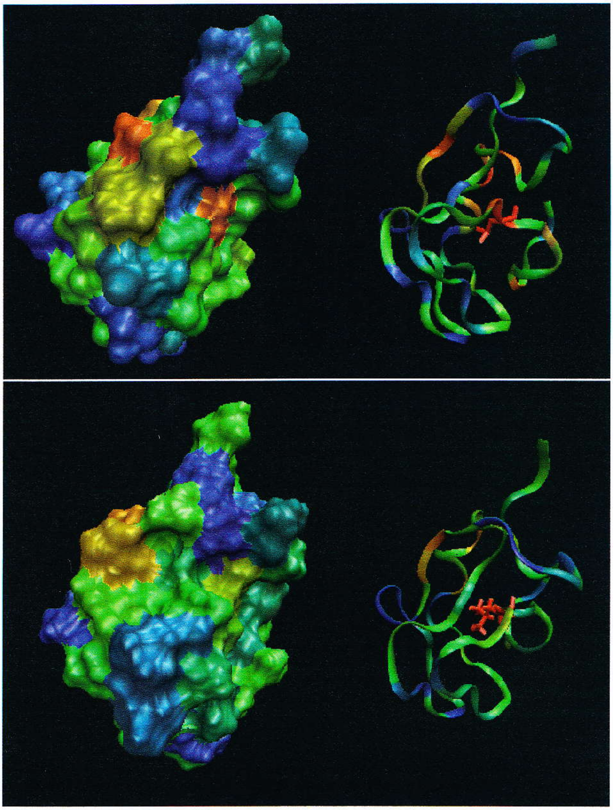


Figure 10. The three-dimensional distribution of $\Delta\tilde{H}_j$ distribution according to color scale shown in Fig.1 and 2. A – the 3-D and ribbon presentation of 1MSI – only one amino acid shown in red – high hydrophobicity deficiency. B –the 3-D and ribbon presentation of 1KDE showing the surface in green (hydrophobicity density accordant to expected one) and blue (higher than expected hydrophobicity density) presumably responsible for preventing the water molecules ordering. The color scale applied as in all other pictures in this chapter.

These proteins support the reliability of the "*fuzzy oil drop*" model being folded almost entirely accordant with the idealized hydrophobicity density distribution (in respect to other proteins demonstrating significant irregularities versus the idealized distribution).

The *SE* scale applied to compare the entropy of the hydrophobicity irregularity distribution calculated for these two proteins (given in Tab.5.) show high similarity of $\Delta\tilde{H}_j$ profiles.

Table. 5. The SE characteristics of antifreeze protein and its mutant.

PROTEIN	SE_+	SE_{+max}	SE_{+REL}	I_+	SE_-	SE_{-max}	SE_{-REL}	I_-
1KDE	2.76	3.17	0.13	32.85	2.26	3.17	0.29	37.64
1MSI	2.82	3.32	0.15	38.80	2.59	3.46	0.25	35.63

It is expected to find other molecules the biological function of which allows the structure accordant with the idealized "*fuzzy-oil-drop*"-like distribution of the hydrophobicity density.

Identification of mutation effects

The important issue in biochemical research is the mutation influence on the structure and biological function. The good example is the Shiga toxin, which is present in PDB as WT (1DM0) [30] and few mutations (for example 1C48 [31]). The *SE* characteristics of these proteins is given in Tab.6.

The comparable analysis of $\Delta\tilde{H}_j$ profiles of WT and mutant allows localize the structural and probably functional changes introduced by mutation (Fig.11.). The long range effects are of particular interest. The comparable analysis and quantitative measurements of mutation effects may be also expressed using *SE* parameters given in Tab.6.

Table 6. The SE characteristics describing different mutants of Shiga toxin. The influence of complexation (denoted by "C") may also be observed and analyzed. The values averaged for 10 sub-units in complex. The values in bold describe the proteins discussed in details below.

	Shiga toxin								
PROTEIN CHAIN	Mutation	SE_+	SE_{+MAX}	SE_{+REL}	I_+	SE_-	SE_{-MAX}	SE_{-REL}	I_-
1DM0 B-K	WT	**2.37**	**3.32**	**0.28**	**29.49**	**2.64**	**3.32**	**0.21**	**36.90**
1C48 A-E	G62T	**2.71**	**3.46**	**0.26**	**35.26**	**2.72**	**3.46**	**0.21**	**34.66**
1C4Q A-E	F30A/W34A"C"	2.51	3.46	0.27	42.06	2.69	3.46	0.22	28.71
1CQF A-E	"C"	2.38	3.46	0.31	30.98	2.66	3.32	0.20	34.39
1CZG A-E	G62T	2.39	3.17	0.29	30.28	2.32	3.17	0.26	30.93
1CZW A-J	W34A	2.52	3.46	0.27	42.02	2.74	3.46	0.21	34.18
1D1K A-E	D17E/W34A"C"	2.48	3.46	0.28	36.04	2.73	3.46	0.21	34.49

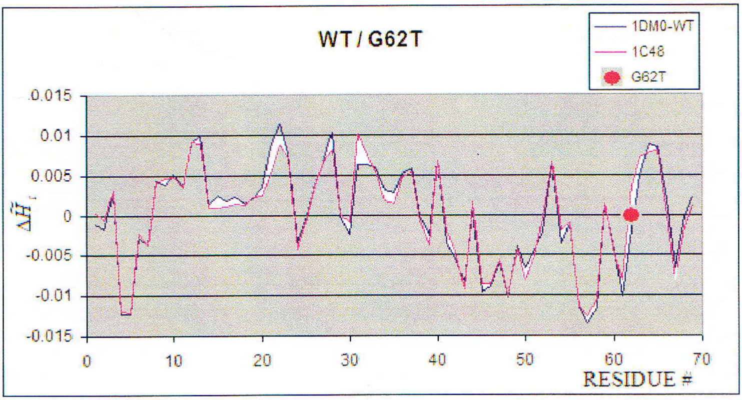


Figure 11. The effects of mutation (G62T) influencing the $\Delta\tilde{H}_j$ profile. The white area distinguishes the differences between the WT and mutated form of protein. The red dot points the mutation position.

Conclusions

The active site recognition *in silico* is of high importance nowadays when many proteins of unknown biological function, identified as products of genome analysis await for a unified automated method allowing recognition of their biological activity [32]. The next step is to develop methods able to predict protein's function from an examination of its structure. Some of the techniques used to identify functionally important residues from the sequence or structure are based on searching for homologues of proteins of known function [33,34]. Homology based technique for proteins of the sequence similarity below 25% fail to identify the biological activity [35]. The stabilization of tertiary structure seems to be based on the hydrophobic interaction [36-38]. The model of hydrophobicity hidden in the central part of protein with hydrophilic residues exposed on the surface of protein is commonly accepted [39-41]. The nonrandom distribution of hydrophobicity was examined in details [42]. Detailed analyses of the spatial variation of hydrophobicity focused on the region of transition between the protein interior and exterior were carried out for 30 relatively diverse globular proteins and for 14 decoys [43]. The transmembrane proteins due to their specificity to be anchored in hydrophobic environment represent high specificity in respect to the hydrophobicity distribution [44]. The hydrophobic interaction was recognized as the main driving force for protein folding process [45,46]. A complex analysis of protein interfaces and their characteristics versus highly divergent areas is presented in [47]. The detailed analysis of the surface shape (geometric irregularities and cavities) is presented in [48]. The $\Delta\tilde{H}_j$ profiles are shown to be the tool for biological activity recognition. The examples given in this chapter present the proteins in which the ligand binding cavity may be recognized analyzing the $\Delta\tilde{H}_j$ maxima. The hydrophobicity deficiency which is probably generated by the empty space (cavity) may be driving force for compatible ligand (with compatible hydrophobicity density distribution over the ligand molecule) to find the proper orientation and localization to generate the specific complex. It is expected that only part of proteins complexed with ligands satisfy the conditions of this model, although even small number of proteins following such mechanism seems to be satisfactory for "*fuzzy-oil-drop*" model.

The antifreeze proteins recognized as molecules following the strategy of "*fuzzy-oil-drop*" model make this model acceptable and probably reliable.

The entropy scale introduced to classify and compare different proteins – their structure and probably also their function – may not be applicable for all proteins. However the specificity of hydrophobicity deficiency/excess distribution all over the protein molecule may be treated as characteristic for particular protein molecule and thus may be used for comparable analysis.

Other techniques oriented on active site identification based mostly on geometric analysis and similarity of ligand binding sites are available in Internet SARIG 65 [49], Q-site Finder [5], Hippo [51], Sprout68 [52], Feature 69-71 [53], WebFEATURE [54], [55], Thematics [56], [57], [58], Apropos [59], Drugsite [60], Ligsite [61], Sumo [62,63], Profunc [64,65].

The large spectrum of proteins representing mostly enzymes the biological function of which was recognized on the basis of "*fuzzy-oil-drop*" is presented in [66]. The CSA data base has been taken as the golden standard [67]. The predictability of the biological function on the basis of "*fuzzy-oil-drop*" model was compared with result of other methods (*Sumo* and *ProFunc*) revealing quite high accordance of these methods [66]. The complexation of protein to ligands and to other protein molecules seems to work satisfactory also taking the $\Delta\tilde{H}_j$ profiles as the criterion for complexation.

The $\Delta\tilde{H}_j$ profiles may be used to identify the effects of mutation. The comparison of proteins of WT and mutants or comparison of two different mutants together with the SE scale may successfully describe particularly the long range effects on structure/function.

The general conclusion taken from the analysis presented in this chapter suggests that the specificity of binding cavity which must be present to ensure the proper complexation of ligand to target protein is that the ligand probably shall be present in the protein folding environment. One may say even more: the ligand seems to be necessary as the active participant in protein folding process influencing it in the aim-oriented form. The ligand with its own hydrophobicity characteristics may occupy the appropriate position in "*fuzzy-oil-drop*" signaling the folding protein the necessity to generate the specific cavity.

The server for active site recognition according to the presented model is available on webpage: www.activesite.cm-uj.krakow.pl.

The simulation of protein folding in the presence of specific ligand will be shown in the next chapter.

References

1. Kauzmann, W.1959. Adv Protein Chem, 14, 1-63.
2. Levitt M. 1976 J Mol Biol, 104(1). 59-107.
3. Weaver L.H, Grutter M.G, Matthews B.W 1995 J. Mol. Biol. 245, 54-68.
4. Vojtechovsky J, Chu K, Berendzen J, Sweet R.M, Schlichting I 1999. Biophys J. 77, 2153-2174.
5. Kragelund B.B, Andersen K.V, Madsen J.C, Knudsen J, Poulsen F.M. 1993. J. Mol. Biol. 230, 1260-1277.
6. 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, Bode W. 1998. Proc Natl Acad Sci USA 95, 3408-3412.
7. Song H, Inaka K, Maenaka K, Matsushima M. 1994. J. Mol. Biol. 244, 522-540.
8. Aguilar C.F, Thomas P.J, Mills A, Moss D.S, Palmer R.A. 1992. J. Mol. Biol. 224, 265-267.
9. Warren M.S, Brown K.A, Farnum M.F, Howell E.E, Kraut J 1991 Biochemistry 30, 11092-11103.
10. Wang Z, Canagarajah B.J, Boehm J.C, Kassisa S, Cobb M.H, Young P.R, Abdel-Meguid S, Adams J.L, Goldsmith E.J. 1998. Structure 6, 1117-1128.
11. 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, Laue E.D. 1998. Nature 395, 244-250.
12. Bossemeyer D, Engh R.A, Kinzel V, Ponstingl H, Huber R. 1993. EMBO J. 12, 849-859
13. Avris C.E, Boyle F.T, Calvest A.H, Curtin N.J, Endicott J.A, Garmon E.F, Gibson A.E, Golding B.T, Grant S, Griffin R.J, Jewsbury P, Johnson L.N, Lawrie A.M, Newell D.R, Noble M.E, Sausville E.A, Schultz R, Yu W. 2000. J. Med.Chem. 43, 2797-2800

14. Nagar B, Bornmann W, Pellicena P, Schindler T, Veach D.R, Miller W.T, Clarkson B, Kuriyan J. 2002 Cancer Res. 62, 4236-4243
15. Liao D.I, Kapadin G, Ahmed H, Vasta G.R, Herzberg O. 1994 Proc Natl Acad Sci USA 91, 1428-1432.
16. http://scop.mrc-lmb.cam.ac.uk/scop).
17. Brylinski M, Konieczny L, Roterman I. 2007. Int. J. Bioinformatics Research and Applications. 3, 234-260.
18. Brylinski M, Konieczny L, Roterman I. 2006. Bioinformation 1, 127-129
19. Shannon, C.E.A.1948. Bell Syst Tech J, 27, 379-423.
20. Brylinski M, Kochanczyk M, Broniatowska E, Roterman I. 2007. J. Mol. Model 13, 665-675.
21. Kanagawa M, Yokoyama S, Kuramitsu S. – to be Publisher
22. Lunin V.V, Skarina T, Onopriyenko O, Binkowski T.A, Joachimiak E, Edwards A.M, Savchenko A. 2008 – to be Publisher
23. Forouhar F, Su M, Jayaraman S, Conover K, Ciao R, Acton T.B, Montelione G.T, Hunt J.F, Tong L, - to be Publisher
24. Ruhul Momen A.Z.M, Sirota H, Hayashi F, Yokoyama S. – to be published
25. Ohnishi S, Kigawa T, Saito K, Kosiba S, Inoue M, Yokoyama S. – to be published
26. Vaynberg J, Fukuda T, Chen K, Vinogradova O, Velyvis A, Tu Y, Ng L, Wu C, Qin J. 2005 Mol. Cell 17, 513-523.
27. Prymula K, Roterman I. 2008 – submitted
28. Jia Z, DeLuca C.I, Chao H, Davies P.L. 1996 Nature 384, 285-288.
29. Sonnichsen F.D, DeLuca C.I, Davies P.L, Sykes B.D. 1996 Structure 4, 1325-1337.
30. Schumacher M.A, Scott D.M, Mathews I.I, Ealick S.E, Roos D.S, Ullman B, Brennan R.G. 2000 J. Mol. Biol. 298, 875-893
31. Ling H, Brunton J.L, Read R.J. – to be published
32. Burley SK. 1999, Nat Genet. 23, 151-157.
33. Bork P J. 1998, Mol. Biol. 283, 707-725.
34. Skolnick J, Fetrow J.S. 2000, Trends Biotechnol. 18, 34-39.
35. Devos D, Valencia A. 2000, Proteins 41, 98-107.
36. Klapper M.H. 1971 Biochim Biophys Acta 229, 557-566.
37. Klotz I.M. 1970, Arch Biochem Biophys 138, 704-706.
38. Meirovitch H, Scheraga H.A. 1980 Macromolecules 13, 1398-1405.
39. Kyte J, Doolottle R.F. J. 1982 Mol. Biol. 157, 105-132.
40. Meirovitch H, Scheraga H.A. 1981 Macromolecules 14, 340-345.
41. Rose G.D. Roy S. 1980 Proc Natl Acad Sci USA 77, 4643-4647.
42. Irbäck A, Peterson C, Potthast F. 1996, Proc Natl Acad Sci USA 93, 9533-9538.
43. Silverman BD 2001, Proc Natl Acad Sci USA 98, 4996-5001.
44. Silverman BD 2003, Protein Sci 12, 586-599.
45. Baldwin RL 2002, Science 295, 1657-1658.
46. Finney JL, Bowron DT, Daniel RM, Timmins PA, Roberts MA 2003, Biophys Chem 105, 391-409.
47. Jimenez JL 2005, Proteins 59, 757-764.
48. Lei H, Duan Y 2004, Protein Eng Des Sel 17, 837-845.

49. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D. 2004, J. Mol. Biol. 344, 1135-1146.
50. Laurie AT, Jackson RM. 2005. Bioinformatics. 21, 1908-1916.
51. Gillet VJ, Myatt G, Zsoldos Z, Johnson AP 1995, Perspect Drug Discov Design. 3, 34-50.
52. Law JMS, Funk DYK, Zsaldos Z, Simon A, Szabo Z. 2003, J. Mol. Struct THEOCHEM 651-657, 666-667.
53. Wei L, Altman RB. 1998, Pac Symp Biocomp 497-508.
54. Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB.2003, Nucleic Acids Res. 31, 3324-3327.
55. Bantao DR, Altman RB, Klein TE 2003, Nucleic Acids Res. 31, 4450-4460.
56. Ko J, Murga LF, Wei Y, Ondrechen MJ 2005, Bioinfomatics 21, (supl. 1) i258-265 2005.
57. Shehadi IA, Abyzov A, Uzun A, Wei Y, Murga LF 2005, J. Bioiform Comput Biol 3, 127-143.
58. Ko J, Murga LF, Andre P, Yang H, Ondrechen MJ, 2005. Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. Proteins 59, 183-195.
59. Peters KP, Fauk J, Frommel C. 1996. J. Mol. Biol. 256, 201-213.
60. An J, Totrov M, Abagyan R. 2004. Genome Inform 15, 31-41.
61. Hendlich M, Rippman F, Barnickel G. 1997 J. Mol. Graph Model. 15, 359-363.
62. Jambon M, Imberty A, Deléage G, Delfaud F. 2006. Bioinformatics 21, 3929-3930.
63. Jambon M, Andrieu O, Combet C, Deléage G, Geourjon C. 2003 Proteins 52, 137-145.
64. Laskowski RA, Watson JD, Thornton JM. 2005. J. Mol. Boil. 351, 614-626.
65. Laskowski RA, Watson JD, Thornton JM. ProFunc: 2005. Nucleic Acids Res 33, W89-W93.
66. Brylinski M, Prymula K, Jurkowski W, Kochanczyk M, Stawowczyk E, Konieczny L, Roterman I. 2007 PLoS Computational Biology 3 (4), e94
67. Porter C.T, Bartlett G.J, Thornton J.M, 2004, Nucleic Acid Res. 8, 3-7.