"Fuzzy oil drop" model applied to individual small proteins built of 70 amino acids

Katarzyna Prymula, Kinga Salapa, Irena Roterman

Editorial Manager(tm) for Journal of Molecular Modeling
Manuscript Draft

Manuscript Number: JMMO1009R1

Title: "Fuzzy Oil Drop" Model Applied to Individual Small Proteins Built of 70 Amino Acids

Article Type: Original paper

Keywords: Biological activity, Active site recognition, Hydrophobicity deficiency, Small proteins, Proteins of unknown function

Corresponding Author: Professor Irena Roterman, Dr

Corresponding Author's Institution: Collegium Medicum - Jagiellonian University

First Author: Katarzyna Prymula

Order of Authors: Katarzyna Prymula; Kinga Salapa; Irena Roterman, Dr

Abstract: The proteins composed of short polypeptides (about 70 amino acid residues) representing the following functional groups (according to PDB notation): growth hormones, serine protease inhibitors, antifreeze proteins, chaperones and proteins of unknown function, were selected for structural and functional analysis. Classification based on the distribution of hydrophobicity in terms of deficiency/excess as the measure of structural and functional specificity is presented. The experimentally observed distribution of hydrophobicity in the protein body is compared to the idealized one expressed by a three-dimensional Gauss function. The differences between these two distributions reveal the specificity of structural/functional characteristics of the protein. The residues of hydrophobicity deficiency versus the idealized distribution are assumed to indicate cavities with the potential to bind ligands, while the residues of hydrophobicity excess are interpreted as potentially participating in protein-protein complexation. The distribution of hydrophobicity irregularity seems to be specific for particular structures and functions of proteins. A comparative analysis of such profiles is carried out to identify the potential biological activity of proteins of unknown function.

Response to Reviewers: Comments to the reviewers

Reviewer #1

The sentence "protein structure determines its biological function" got changed to the form :
"The spatial distribution of amino acid residues and particularly distribution of their specific hydrophobicity in a protein structure is assumed to influence the biological function".

Reviewer #2

Ad.1. The section INTRODUCTION has been modified making the problem of NBP less exposed. The paper describing the structures of NBP form is in press currently. The appearance of that paper will make clear the idea of NBP and the usefulness of the presented method. The real proteins available in PDB allowed the comparative analysis of the NBP proteins reveling some substantial differences and similarities between these two groups of proteins position #18 on the reference list (Prymula K, Piwowar M, Kochanczyk M, Flis L, Malawski M, Szepieniec T, Evangelista G, Minervini G, Polticelli F,

Wisniowski Z, Salapa K, Matczynska E, and Roterman I (2009) In silico structural study of random amino acid sequence proteins not present in nature. Chem Biodivers, in press).

Ad.2. The hydrophobicity scale
The hydrophobicity scales (theoretical and experimental) were compared in details in respect to "fuzzy oil drop" model in other publication. The "fuzzy oil drop" estimates the hydrophobicity distribution in the very simplified and averaged form. In result, the relative distribution of hydrophobicity is under consideration. The differences observed between different scales get marginal in this strongly averaged form. The expression says rather that any hydrophobicity scale may be applied to the model depending on the users preferences.

Ad.3. The expression "in aim-oriented form" got changed to the form:
"The ligand binding sites, ion binding sites, protein-protein interactions area can be recognized on the basis of the hydrophobicity excess/deficiency distribution all over the protein body [13-17, 19-24]".

Ad.4. The correlation coefficients were recalculated taking into the base all proteins presented in the paper. Although the program calculating the sequence similarity did not accepted all pairs to be compared. This is why the final number of protein pairs appeared to be 180.
The new correlation coefficients and new pictures are given in Fig.7. The appropriate discussion is added.

Ad.5. The text was corrected (particularly in respect to the repetitions concerning NBP).

"Fuzzy oil drop" model applied to individual small proteins built of 70 amino acids

Received: 05.09.2009 / **Accepted**: 16.12.2009

.

Katarzyna Prymula[1,2], Kinga Sałapa[1], Irena Roterman[1,✉]

(1) Department of Bioinformatics and Telemedicine Jagiellonian University – Collegium Medicum, Lazarza 16, 31-530 Krakow, Poland

(2) Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Krakow, Poland

✉Tel/Fax: +48 12 619 96 93; Email: myroterm@cyf-kr.edu.pl

Abstract

The proteins composed of short polypeptides (about 70 amino acid residues) representing the following functional groups (according to PDB notation): growth hormones, serine protease inhibitors, antifreeze proteins, chaperones and proteins of unknown function, were selected for structural and functional analysis. Classification based on the distribution of hydrophobicity in terms of deficiency/excess as the measure of structural and functional specificity is presented. The experimentally observed distribution of hydrophobicity in the protein body is compared to the idealized one expressed by a three-dimensional Gauss function. The differences between these two distributions reveal the specificity of structural/functional characteristics of the protein. The residues of hydrophobicity deficiency versus the idealized distribution are assumed to indicate cavities with the potential to bind ligands, while the residues of hydrophobicity excess are interpreted as potentially participating in protein-protein complexation. The distribution of hydrophobicity irregularity seems to be specific for particular structures and functions of proteins. A comparative analysis of such profiles is carried out to identify the potential biological activity of proteins of unknown function.

Keywords Biological activity · Active site recognition · Hydrophobicity deficiency · Small proteins · Proteins of unknown function

Introduction

Techniques aimed at engineering proteins for pharmacological use, with enhanced stability or improved functions, are being widely developed nowadays [1-3]. The investigation of proteins exhibiting desirable activity among the proteins present in organisms, as well as the generation of new proteins resulted in the development of polypeptide sequence libraries [4, 5]. Designing polypeptides that modify the biological activity of other proteins [6], exhibit new or altered catalytic properties [7] and high stability [8] is one of the goals of pharmaceutical research. The calculation of the number of all possible sequences containing 50 amino acid residues composed of 20 standard amino acids revealed that only a minority of these sequences occur in nature [9]. The huge number of protein sequences not observed in nature seems to be the space hiding many biological functions not exploited in the biochemical pathways developed so far by living organisms. Attempts focused on the investigation of protein folds within libraries containing totally random sequences have been made [9, 10]. The search for pharmacologically active proteins among Never Born Proteins (NBPs) is one of the aims of the EUChinaGRID project [11]. In the project, the search for biologically active proteins that have not been created during evolution is performed *in silico* using two complementary methods: ROSETTA [12] and the "fuzzy oil drop" (FOD) model [13-17]. The results of this investigation are presented elsewhere [18].

Before the space of NBPs can be exploited, the existing proteins of defined chain length were taken under consideration with respect to their active-site properties. The characterization of biological activity of known proteins can help to determine the possible biological functions of NBPs. In our study, the position of potential active sites (understood also as the ligand-binding sites) is assigned on the basis of hydrophobicity deficiency in particular proteins. The biological activity has been identified based on the FOD model [19-22]. The number of sequences having 70 amino acid residues and a known 3-D structure is relatively large. It is impossible to present their characteristics in one paper. Proteins participating in the formation of large complexes (ribosomes) and those interacting with DNA or RNA as well as proteins with bound ligands are described elsewhere [23, 24].

Materials and methods

Data

The tool available on the Protein Data Bank [25] webpage aimed at searching for proteins that satisfy particular criteria was used to select proteins according to the defined polypeptide chain length (70 amino acid residues). Proteins containing less than 72 and more than 68 amino acids were pre-selected from the group of entries found in PDB (July 2007). In the next step, the structures which lacked the coordinates of main chain atoms in the middle of the sequence were rejected. The proteins chosen for analysis and presented here are a subset of the selected entries and belong to one of the following functional groups: growth factors, serine protease inhibitors, antifreeze proteins, chaperones and proteins of unknown function (Table 1). The data for the analyzed structures were taken from investigations carried out using X-ray crystallography [26-28] and NMR techniques [29-40]. Additionally, the SH3 domain is discussed with respect to the various proteins that contain it and their different biological functions [41].

Sequence and structure analysis

Protein sequences and structures were compared within the set of selected chains (Table 2). Sequence alignment was performed using the LALIGN program [42] from the FASTA package [43], version 35, with scoring matrix PAM250 and gap open/extension penalties equal to -10/-2. Additionally, for sequences from the SH3 domain family multiple sequence alignment was produced with the aid of ClustalW [44]. Structure similarity was measured using DaliLite [45] for pairwise structure comparison and resulted RMS-D values calculated for the C_α atoms were taken for further analysis (see the Statistical section for details).

Binding site recognition

The FOD model was applied to identify the distribution of hydrophobicity in the selected proteins [17, 21]. The idealized hydrophobicity distribution is represented by the three dimensional Gauss function, according to Eq. 1:

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

$\tilde{Ht}_j$ represents the theoretical hydrophobicity at *j*-th point (x_j, y_j, z_j) – position of the effective atom (averaged side chain position or C_α atom for GLY), while the inverse of $\tilde{Ht}_{sum}$ is the normalizing constant. The point at which the Gauss function reaches its maximum is located in the geometric center of the macromolecule. Theoretical hydrophobicity is calculated after the translation of a molecule, so that its centroid coincides with the origin (this is why all mean values are equal to 0). The values of $\sigma_x, \sigma_y, \sigma_z$ (traditionally interpreted as standard deviation), calculated separately for each dimension, represent the size of the "drop", which depends on the length of the polypeptide chain (and the size of a macromolecule) under consideration [14].

The empirical hydrophobicity distribution (as observed based on the hydrophobic properties of residues in real proteins) is calculated by applying the Levitt [46] function, according to Eq. 2:

$$\tilde{Ho}_j = \frac{1}{\tilde{Ho}_{sum}} \sum_i \tilde{H}_i^r + \tilde{H}_j^r \begin{cases} \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right] & \text{for } r_{ij} \le c \\ 0 \text{ for } r_{ij} > c \end{cases} \quad (2)$$

$\tilde{Ho}_j$ represents the empirical hydrophobicity of the *j*-th amino acid residue; $\tilde{H}_i^r$, $\tilde{H}_j^r$ are the hydrophobicity characteristics for the *i*-th and *j*-th amino acid residues, respectively; r_{ij} is the distance between the *j*-th and *i*-th effective atoms representing the side chain position of a particular amino acid; *c* expresses the cut-off distance, which has a fixed value equal to 9.0 Å, as stated in the original paper [46]; finally, the inverse of $\tilde{Ho}_{sum}$ is the normalizing constant. $\tilde{Ho}_j$ collects the hydrophobic interactions of the *j*-th residue with others localized within the specified distance *c*. Any hydrophobicity scale [47-53] may be applied to calculate the observed distribution of hydrophobicity. Despite some differences between hydrophobicity parameters in these scales, it was found that these differences do not significantly influence the final results [18]. The hydrophobicity scale applied for the calculations presented in this

paper is based on the "fuzzy oil drop" model. The hydrophobicity parameter is estimated according to the relative position in the three-dimensional Gauss function. This approach makes possible the calculation of the hydrophobicity of ligand molecules. Since the model ("fuzzy oil drop" model) is assumed to be used for folding process simulation in the presence of ligand molecule [15, 17] the self-consistency of the hydrophobicity scales for amino acids and ligand molecules is of high importance [17]. The quantitative comparison of this scale with commonly used ones is also shown in [15].

Since both values of hydrophobicity (theoretical and empirical) are normalized, the difference between these values, calculated according to Eq. 3, represents the deviation of the empirical hydrophobicity versus the theoretical one, at a particular point in space (effective atom of the *j-th* amino acid residue).

$$\Delta\tilde{H}_j = \tilde{H}t_j - \tilde{H}o_j \qquad (3)$$

$\Delta\tilde{H}_j$ maxima are related to hydrophobicity deficiency, which is expected to indicate a potential binding site. The potential ligand may bind in this area as a complementary element compensating the hydrophobicity deficiency and producing a smoothed hydrophobicity distribution. Negative values of $\Delta\tilde{H}_j$ represent an area of excess hydrophobicity. When located on the protein surface, an area with such characteristics is expected to represent a potential area responsible for protein-protein interactions.

The similarity estimation on the basis of the $\Delta\tilde{H}_j$ profile

The spatial distribution of amino acid residues and particularly distribution of their specific hydrophobicity in a protein structure is assumed to influence the biological function. The $\Delta\tilde{H}_j$ profiles also appeared sequence- and structure-dependent. Consequently, the $\Delta\tilde{H}_j$ profiles expressing the irregularities (deficiency or excess) of hydrophobicity are used to compare and measure the specificity of particular proteins. The $\Delta\tilde{H}_j$ distribution in the protein structure, represented for example in the form of ribbon model, reveals a remarkable discrepancy of hydrophobicity (deficiency/excess) between clusters of residues, which may be treated as potential functional sites.

Information entropy

Screening the hydrophobicity discrepancy along the polypeptide chain (presented in the form of a $\Delta\tilde{H}_j$ profile) reveals fragments of sequence characterized by high values of $\Delta\tilde{H}_j$. Two extreme instances of hydrophobicity discrepancy profile may be distinguished. The first one occurs when there is exactly one well-defined sequence fragment (consecutive residues) with high positive $\Delta\tilde{H}_j$ values, suggesting that this fragment forms a binding site. The second case takes place when there is more than one sequence fragment with positive $\Delta\tilde{H}_j$, evenly distributed along the polypeptide chain. The degree of predictability of the formation of such functional sites is high in the first case, while the second case represents low predictability. The level of predictability can be easily measured using the information theory – and information entropy in particular. The amount of information carried by a particular system (represented by $\Delta\tilde{H}_j$ profile in our example) may be calculated.

The first case can be treated as deterministic (residues of one fragment with high $\Delta\tilde{H}_j$ are determined to meet in the space). The second case represents a random case (with randomness understood as equal probability for each fragment to participate in a common structural element).

The case with many $\Delta\tilde{H}_j$ maxima can also represent a protein with one or more binding sites. However, small polypeptides (around 70 amino acid residues) are expected to contain at best one well-defined binding site.

The characterization of a binding site can be expressed quantitatively using the probability theory.

The entropy of a binding site involving residues close in sequence is low (first case), compared with the entropy of a binding site formed by residues evenly distributed in sequence (second case). Information entropy (*SE*) calculated for fragments with positive $\Delta\tilde{H}_j$ ($\Delta\tilde{H}_j^p$) according to Eq. 4 measures the amount of predictability (according to original definition – uncertainty) of the organization of residues forming a binding site (see the examples discussed in [20]):

$$SE_+ = -\sum_j^K p_j \log_2 p_j \qquad (4)$$

where K denotes the number of fragments with positive $\Delta\tilde{H}_j$, and

$$p_j = \sum_{i=1}^{N_{ij}} \frac{\Delta\tilde{H}_i^p}{\Delta\tilde{H}_t^p} \tag{5}$$

where N_{ij} represents the number of positive $\Delta\tilde{H}_j$ values belonging to j-th fragment and $\Delta\tilde{H}_t^p$ is the sum of all positive $\Delta\tilde{H}_j$ values of the whole polypeptide chain.

The SE_+ characterizes a particular protein and may describe an active site (fragments with positive values). Another parameter used in our study of binding sites is the information necessary to localize residues creating a binding site (I [bit]). The participation of particular residues in active site creation is understood as the probability of the conjunction of events (close mutual localization) and can be calculated according to Eq. 6:

$$I = -\log_2 \prod_{j=1}^{K} p_j \tag{6}$$

where K has the same meaning as in Eq. 4.

Both parameters (SE, I) describe similar characteristics of the active site formed by selected residues participating in the organization of a particular function-related structure

The $SE^{\max}$ (the highest uncertainty) describes the situation when all solutions are equally probable (all positive $\Delta\tilde{H}_j$ fragments are equally probable – discussed as the second case).

$$SE_+^{\max} = \sum_{j=1}^{K} \frac{1}{K} \log_2 \frac{1}{K} \tag{7}$$

The $SE^{\max}$ calculated for equal p_j values describes the random situation in which each fragment is equally represented (in the scale of p). The difference between $SE^{\max}$ and SE and may be used to measure the distance in the probability between a particular solution (active site construction) and a completely random result [20].

The same analysis may be applied to negative $\Delta\tilde{H}_j$ values (SE_-).

Statistical analysis

The parameters describing the similarity of proteins under consideration were analyzed from a statistical point of view. The correlation coefficient measuring the relation between sequence similarity (expressed by Waterman-Eggert score) and structural similarity (expressed by RMS-D), and biological function similarity (expressed by differences between *SE* and *I* parameters introduced in section 2.5, and shown in Table 2) was calculated using the *Statistica* Program [54]. All pairs of chains listed in Table 2 were taken into account in this analysis. The level of significance was selected at α=0.05. The Spearman's correlation coefficient [55] was calculated due to the absence of normal distribution. It is a non-parametric measure of correlation between two variables, without making any other assumptions about the particular nature of the relationship between the variables and no preliminary distribution analysis (for normal distribution of the variables the calculation of the Pearson correlation coefficient is applicable).

The correlation coefficient was calculated also pair-wise for proteins representing similar $\Delta\tilde{H}_j$ profiles to measure the degree of similarity between them.

Results and discussion

The set of proteins fulfilling the polypeptide chain length criterion (approx. 70 amino acid residues) is quite large and diverse with regard to their biological function. The analysis of proteins crystallized as individual molecules is presented in this paper.

$\Delta\tilde{H}_j$ profile analysis

All proteins in the presented set are characterized based on $\Delta\tilde{H}_j$ profiles and *SE* calculations. The similarity between $\Delta\tilde{H}_j$ profiles is analyzed in respect to structural similarity and to the results of sequence alignment. The same interpretation is assumed in respect to the *SE* values. If the *SE* based comparison appears compatible with structure and sequence comparison it could be used as equivalent to the other methods requiring the superposition or alignment.

Growth factors

The group of insulin-like growth factors I (IGF-1) represented in PDB by four entries (Table 1) which appeared to be of identical sequence. This is why solely 1BQT is presented in this paper. The $\Delta\tilde{H}_j$ profile shown in Fig. 1a together with the spatial distribution of $\Delta\tilde{H}_j$ values (Fig. 1b,c) is quantitatively characterized in Table 2. Other proteins belonging to this group represent some differences in $\Delta\tilde{H}_j$ profile although they are not of significant importance (results not shown here). Thorough analysis, however, reveals that the $\Delta\tilde{H}_j$ profile of 1PMX differs the most from the others. This observation coincides with the fact that only 1PMX exhibits the structure of IGF-1 complexed with a peptide inhibitor.

Both the structure of IGF-1 complexed with the peptide identified from the phage display library [35], and the structure of the same protein complexed with a detergent [56] revealed that the binding sites are located in fragments B (residues 1-29) and A (residues 42-62). The interactions between IGF-1 and peptides are hydrophobic and formed by residues 3,4,13,17 and 54. The $\Delta\tilde{H}_j$ values of these resides represent local maxima or are in their close neighborhood (Fig. 1b, c).

The correlation coefficients calculated pair-wise for $\Delta\tilde{H}_j$ profiles of proteins that belong to the group of growth factors appeared to be significant (very low values of p). It suggests that the $\Delta\tilde{H}_j$ profile similarity is able to represent the structural (functional) similarity of proteins under consideration.

Serine protease inhibitors

The comparison of the $\Delta\tilde{H}_j$ profiles (Fig. 2a) as well as the *SE* values (Table 2) for proteins representing serine protease inhibitors indicates that although these proteins share similar function, their structures display some differences.

The $\Delta\tilde{H}_j$ profiles (Fig. 2a) and the structure representation (Fig. 2b) visualize to what extent the $\Delta\tilde{H}_j$ similarity influences the values of *SE* parameters.

The local $\Delta\tilde{H}_j$ minimum close to residue 15 significantly suggests a common localization of the area potentially responsible for the interaction with other proteins (Fig. 2a), which was

actually confirmed experimentally when the complex structure of the inhibitor with bovine trypsin was resolved [57]. The following residues form a protein-protein interface: 11, 12, 14-18, 20, 21, 23, 52 and 57. The entropy of such a functional site should therefore be small, and this is confirmed by the *SE*- values (Table 2), which are lower than those for the overwhelming majority of structures analyzed in this study. The $\Delta\tilde{H}_j$ values of these residues do not suggest the recognition of these residues to be responsible for biological function. Neither local maxima nor local minima are represented by these residues.

The correlation coefficient calculated for $\Delta\tilde{H}_j$ values appeared significant only for the pair 1EGL and 1DWM suggesting high similarity between these two proteins. The 1BBI molecule differs versus the others (Fig. 2c) what can also be seen taking S scale as the criterion.

Antifreeze proteins

This functional class is represented by proteins belonging to the group responsible for protection against the ordering of water molecules, and the subsequent formation of ice.

The antifreeze protein from the North Atlantic ocean pout (1MSI) and its mutated counterpart (1KDE), which exhibit 97.14% sequence identity, display very similar (although not identical) $\Delta\tilde{H}_j$ profiles (Fig. 3a) and hydrophobicity distribution (Fig. 3b). In this case, sequence identity is reflected in structure similarity (RMS-D=0.8 Å). The *SE* values are very close to each other (Table 2). The comparison of the *SE* with respect to the sequence/structure similarity may be used to visualize the mutual relation between these two characteristics and shows to what extent the *SE* can be treated as a similarity measure. These examples were selected to show how the sequence and structural similarity influences the $\Delta\tilde{H}_j$ profile and the *SE*.

These two proteins are good examples for the FOD model. The three-dimensional Gauss function taken to represent the hydrophobic environment during the protein folding process directs the hydrophilic residues toward the protein surface and the hydrophobic residues toward the center of the molecule in order to form a spherical shape. In the case of these structures it has succeeded.

Antifreeze proteins lower the temperature of ice growth with respect to the bulk of the solvent. This phenomenon, known as thermal hysteresis, is supposed to be the result of the adsorption of antifreeze proteins on the surface of ice [58-62]. The mechanism by which

antifreeze proteins contribute to thermal hysteresis in fish is not completely understood. Some explanations of this phenomenon suggest the reduction of the solubility of the antifreeze proteins in the solution as one of the causes of hysteretic activity [63]. Other studies [26, 64] identify N14, T18 and Q44 as the key residues for antifreeze protein (type III) – ice interaction and reveal the amphipathic character of the ice-binding site. Interaction with ice is of low specificity character. This is why the detailed interpretation of their $\Delta\tilde{H}_j$ values is not the point. The scale for $\Delta\tilde{H}_j$ values was selected intentionally to visualize the high accordance of the protein structure to the theoretical model suggesting the mechanism of the folding process to be accordant with the assumed model.

The antifreeze proteins display very short fragments (only one residue in 1KDE) of high hydrophobicity deficiency. These fragments are almost entirely "buried" in the central part of the molecule (and inaccessible on the protein surface), which suggests these molecules are unlikely to bind ligands (no hydrophilic cavity in this protein). The protein molecule is covered with residues with $\Delta\tilde{H}_j$ close to 0 (Fig. 3b, yellow color) and some surface fragments exhibiting excess hydrophobicity (Fig. 3b, orange), which may indicate the potential location of an area responsible for protein-protein interactions, as observed in other proteins [18]. On the other hand, the presence of a hydrophobicity deficiency area on the surface is in agreement with the experimental observation which suggests that the possible disorder of water molecules prevents the structuring of water molecules during the freezing process. The residues suspected of being responsible for the antifreeze activity [64] belong to the hydrophilic fragment on the surface (Fig. 3c).

The high value of correlation coefficient (Fig. 3a) confirms high similarity of $\Delta\tilde{H}_j$ profiles noted after visual inspection of them.

Chaperones

Proteins responsible for controlling the folding process are represented by four structures although three of them (1U96, 1U97, 1Z2G) have identical sequences (1U96 and 1Z2G were excluded from the discussion). The protein 2GUZ does not exhibit any significant sequence similarity to 1U97. The *SE* parameters for proteins representing chaperones are given in Table 2. The very well-defined and long fragments of hydrophobicity deficiency can be recognized based on low SE_+. The $\Delta\tilde{H}_j$ profiles (Fig. 4a) with a few marked minima, such as those

observed for these structures, are characteristic of compact shapes with long unstructured loops that stick out (Fig. 4b, c).

The correlation coefficient (Fig. 4a) seems to express quantitatively the relatively low similarity of $\Delta\tilde{H}_j$ profiles for proteins in this group.

The lack of accordance between SE parameters and between $\Delta\tilde{H}_j$ profiles (low value of correlation coefficient) despite of quite high structural similarity seems to be due to the long extended N- and C-terminal polypeptide fragments in 1U97 which are absent in 2GUZ (Fig. 4b,c). This makes the approximation of 1U97 to the sphere (ellipsoid described by 3-D Gauss function) not appropriate.

Proteins of unknown biological function

The proteins belonging to this group are of particular interest. The comparison of $\Delta\tilde{H}_j$ profiles suggests that there are no proteins of mutual similarity in this group, a result confirmed by the *SE* parameter values (Table 2) The comparison of profiles for the proteins of this group with those of proteins of known biological function may be the way to recognize similar proteins and, possibly, their biological function. This is a difficult challenge for the FOD model.

Taking into account all the selected structures containing 70 amino acid residues in the polypeptide chain, some similarities may be found. The $\Delta\tilde{H}_j$ profile of a human protein of unknown function (2CRE) appeared to be similar to the profiles of two structures which belong to the SH3 domain. Namely, these are fragments of the human cytoplasmic protein NCK2 (1U5S:A), classified as a metal-binding protein, and the kinase-binding protein 1 (2DA9) found in mice – a regulator of ubiquitous kinase (Ruk), which regulates apoptosis (Fig. 5a). The ribbon models of these three proteins (2CRE, 1U5S:A and 2DA9) and their surface representations with highlighted areas of hydrophobicity excess/deficiency are shown in Fig. 5b. The sequence similarity between these proteins calculated using the LALING program is summarized in Table 3. The gaps suggested in Fig. 5a appeared to be in agreement with the gap localized using sequence similarity estimation making the 2CRE and 2DA9 $\Delta\tilde{H}_j$ profiles more similar.

The 2CRE and 1U5S:A can be found as proteins of highest similarity in terms of *SE* parameters (Table 2), although higher sequence similarity (according to LALIGN) has been found for 2CRE and 2DA9.

According to the recognition procedure based on $\Delta\tilde{H}_j$ profile analysis, the protein 2CRE of unknown biological function can possibly be a metal-binding protein (as 1U5S-A) or a kinase-binding protein 1 (as 2DA9). Another common characteristic of these three proteins is that they share the SH3 domain fold. The group of SH3 domains will be presented in the next part of this paper.

The comparison of correlation coefficient for $\Delta\tilde{H}_j$ profiles suggests the high similarity between 2CRE and 1U5S:A what is in accordance with the *SE* parameters (Table 2).

The SH3 domains

The SH3 domains constitute a family of proteins characterized by different functions. Nevertheless, they are taken into consideration in the analyzed dataset because of the similarity found between $\Delta\tilde{H}_j$ profiles of one protein with unknown function and two structures from this group (see above). The relatively differentiated $\Delta\tilde{H}_j$ profiles of SH3 domains are shown in Fig. 6a together with the standard deviations (SDs) of $\Delta\tilde{H}_j$ presented in Fig. 6b, calculated for each position of multiple sequence alignment taking into account a non-redundant data set (identical profiles were eliminated). Loops, as well as β strands, contain fragments characterized by high SD (Fig. 6d). Higher variability is, however, observed within areas on fragments with high $\Delta\tilde{H}_j$ (Fig. 6c), therefore suggesting that they coincide with areas of possible biological activity as SH3 domains differ with regard to their function depending on the systems they are part of.

The pair-wise comparison of $\Delta\tilde{H}_j$ profiles by means of correlation coefficient is given in Table 4.

Statistical analysis

In order to quantify to what extent the *SE* and *I* parameters are able to measure the structural/functional similarity the correlation coefficients between parameters describing similarity of sequence, 3-D structure and $\Delta\tilde{H}_j$ profiles were calculated. Sequence alignment scores (Waterman-Eggert), RMS-D values and differences of *SE* and *I* parameters shown in Table 2 were variables used in correlation analysis.

The calculation of correlation coefficients was used to describe relationship between measures of sequence, 3-D structure and $\Delta\tilde{H}_j$ profile similarities. The results are summarized in Table 5. For undisputable similar structures low RMS-D and high Waterman-Eggert score was obtained and negative (as expected) significant correlation between RMS-D and sequence similarity measure was obtained.

The proteins belonging to the category of "unknown function" were different since their structure were not classified according to the function related criteria.

Since no normal distribution has been observed for all variables, the Spearman coefficient was calculated to measure the degree of mutual relation the significance of which was additionally tested. The results are shown in Table 5. The most interesting are the significant correlations between $\Delta\tilde{H}_j$ profile based parameters and those traditionally used for similarity estimation: sequence and RMS-D values. The significant correlation was found to be present between sequence and $|\Delta SE_+|$ and between sequence and $|\Delta SE_-|$. On the other hand the significant correlation between RMS-D (structural similarity) and $|\Delta SE_+^{\max}|$ and between RMS-D and $|\Delta SE_-|$. It suggests that other $\Delta\tilde{H}_j$ based parameters are applicable for sequence and for structure. It may be concluded that the sequence and structure comparison can be performed using $\Delta\tilde{H}_j$ although more extended analysis shall be performed. The technique of multiple alignment (such as multiple sequence alignment – MSA) applicable for the set of $\Delta\tilde{H}_j$ profiles comparison is under consideration. It could make possible large scale comparison of structures introducing the structural insertion/deletion operation.

Conclusions

Prior to the analysis of the biological activity of NBPs, real proteins were characterized with respect to their structural characteristics and active-site architecture. The FOD model was applied to identify the location of possible ligand-binding sites. The ligand binding sites, ion binding sites, protein-protein interactions area can be recognized on the basis of the hydrophobicity excess/deficiency distribution all over the protein body [13-17, 19-24]. The FOD model applied to other proteins (including proteins participating in protein-protein complexes) showed that this model can be used as a tool for active site localization in proteins

(including ligand-binding sites). The applicability and reliability of the FOD model depends on the class of enzymes analyzed, as shown in [22].

The oil-drop model introduced by Kauzman [65] expresses the characteristics of a protein molecule as a construction based on the hydrophobic core in discrete form. The FOD model based on the three-dimensional Gauss function makes this model suitable for the differentiation of protein molecule character. Many proteins presenting irregularities and characteristics deviating from the idealized hydrophobicity distribution were presented elsewhere [15, 17]. The antifreeze proteins which show structures highly in agreement with the idealized model represent very interesting examples of the applicability of the FOD model. The functional characteristics of these proteins, highly soluble in aquatic environments may be seen in Fig. 3. The minimal differences between the theoretical and the observed hydrophobicity throughout almost the entire protein body, as expressed by the $\Delta\tilde{H}_j$ values, which are near zero, confirm the reliability of the model. The irregularity of hydrophobicity observed in many proteins has been the subject of extensive analysis. The relationship between hydrophobicity profiles and biological function, expressed in the hypothesis on the obligatory presence of the ligand during the protein folding process to ensure high specificity towards specific ligands (including substrates), is also currently being explored [66]. The second example of the applicability of the FOD model is the search for similarity. Two SH3 domains – 1U5S:A and 2DA9 – appeared to exhibit *SE* parameters similar to those of 2CRE – a protein of unknown biological function. For this reason, it is possible to put forward a hypothesis that this protein may also have a similar function as the above-mentioned proteins, although such similarity in *SE* parameters should not be taken for granted among SH3 domains. The $\Delta\tilde{H}_j$ profiles seem to be rather function-dependent. For SH3 domains, the location of biological activity (generally represented in red on the $\Delta\tilde{H}_j$ profiles) is in line with the highest variability of $\Delta\tilde{H}_j$ for large fragments with low differentiation, which suggests common structural and different functional characteristics. This observation is understandable given the relatively varied biological activity of the proteins containing SH3 domains.

Acknowledgments

The Authors are very grateful to Prof. Leszek Konieczny (Institute of Medical Biochemistry – Collegium Medicum – Jagiellonian University – Krakow – Poland) for our fruitful discussion. This research was supported by Collegium Medicum grants 501/P/266/L. This study has also been financially supported by the European Commission in the frame of the EUChinaGRID project (contract number: 026634)

Availability

The tools applied for the calculations presented in this paper are available at http://www.bioinformatics.cm-uj.krakow.pl/activesite.

References

1. Rosenberg M, Goldblum A (2006) Computational protein design: a novel path to future protein drugs. Curr Pharm Des 12:3973–3997
2. Deng Y, Zheng Q, Ketas TJ, Moore JP, Lu M (2007) Protein design of a bacterially expressed HIV-1 gp41 fusion inhibitor. Biochemistry 46:4360–4369
3. Antikainen NM, Martin SF (2005) Altering protein specificity:techniques and applications. Bioorg Med Chem 13:2701–2716
4. Patrick WM and Firth AE (2005) Strategies and computational tools for improving randomized protein libraries. Biomol Eng 22:105–112
5. Chaparro-Riggers JF, Polizzi KM, Bommarius AS (2007) Better library design:data-driven protein engineering. J Biotechnol 2:180–191
6. Fowler SB, Poon S, Muff R, Chiti F, Dobson CM, Zurdo J (2005) Rational design of aggregation-resistant bioactive peptides: Reengineering human calcitonin. Proc Natl Acad Sci USA 102:10105–10110
7. Shao Z, Arnold FH (1996) Engineering new functions and altering existing functions. Curr Opin Struct Biol 6:513–518
8. Eijsink VGH, Bjørk A, Gäseidnes S, Sireväg R, Synstad B, van den Burg B, Vriend G (2004) Rational engineering of enzyme stability. J Biotechnol 113:105–120
9. Chiarabelli C, Vrijbloed JW, Thomas RM, Luisi PL (2006) Investigation of de novo totally random biosequences. Part I: A general method for in vitro selection of folded domains from a random polypeptide library displayed on phage. Chem Biodivers 3:827–839
10. Chiarabelli C, Vrijbloed JW, Lucrezia DD, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E, Luisi PL (2006) Investigation of de novo totally random biosequences. Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. Chem Biodivers 3:840–859
11. EUChinaGRID project. http://www.euchinagrid.org/
12. Bonneau R, Strauss CEM, Rohl CA, Chivian D, Bradley P, Malmström L, Robertson T, Baker D (2002) De novo prediction of three-dimensional structures for major protein families. J Mol Biol 322:65–78
13. Konieczny L, Brylinski M, Roterman I (2006) Gauss-function-based model of hydrophobicity density in proteins. In Silico Biol 6:15–22

14. Brylinski M, Konieczny L, Roterman I (2006) Fuzzy-oil-drop hydrophobic force field–a model to represent late-stage folding (in silico) of lysozyme. J Biomol Struct Dyn 23:519–528
15. Brylinski M, Konieczny L, Roterman I (2006) Hydrophobic collapse in (in silico) protein folding. Comput Biol Chem 30:255–267
16. Brylinski M, Konieczny L, Roterman I (2006) Hydrophobic collapse in late-stage folding (in silico) of bovine pancreatic trypsin inhibitor. Biochimie 88:1229–1239
17. Brylinski M, Konieczny L, Roterman I (2007) Is the protein folding an aim-oriented process? Human haemoglobin as example. Int J Bioinform Res Appl 3:234–260
18. Prymula K, Piwowar M, Kochanczyk M, Flis L, Malawski M, Szepieniec T, Evangelista G, Minervini G, Polticelli F, Wisniowski Z, Salapa K, Matczynska E, Roterman I (2009) In silico structural study of random amino acid sequence proteins not present in nature. Chem Biodivers, in press
19. Brylinski M, Konieczny L, Roterman I (2006) Ligation site in proteins recognized in silico. Bioinformation 1:127–129
20. Brylinski M, Kochanczyk M, Konieczny L, Roterman I (2006) Sequence-structure-function relation characterized in silico. In Silico Biol 6:589–600
21. Brylinski M, Kochanczyk M, Broniatowska E, Roterman I (2007) Localization of ligand binding site in proteins identified in silico. J Mol Model 13:665–675
22. Brylinski M, Prymula K, Jurkowski W, Kochańczyk M, Stawowczyk E, Konieczny L, Roterman I (2007) Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput Biol 3:e94
23. Prymula K, Roterman I (2009) Functional characteristics of small proteins (70 amino acid residues) forming protein-nucleic acid complexes. J Biomol Struct Dyn 26:663–677
24. Prymula K, Roterman I (2009) Structural entropy to characterize small proteins (70 aa) and their interactions. Entropy 11:62–84
25. Berman HM, Battistuz T, Bhat TN, Bluhm W F, Bourne P E, Burkhardt K, Feng Z, Gilliland G L, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C (2002) The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 58:899–907
26. Jia Z, DeLuca CI, Chao H, Davies PL (1996) Structural basis for the binding of a globular antifreeze protein to ice. Nature 384:285–288

27. Wisniewska M, Bossenmaier B, Georges G, Hesse F, Dangl M, Künkele KP, Ioannidis I, Huber R, Engh RA (2005) The 1.1 A resolution crystal structure of the p130cas SH3 domain and ramifications for ligand selectivity. J Mol Biol 347:1005–1014
28. Mokranjac D, Bourenkov G, Hell K, Neupert W, Groll M (2006) Structure and function of Tim14 and Tim16, the J and J-like components of the mitochondrial protein import motor. EMBO J 25:4675–4685
29. Werner MH, Wemmer DE (1992) Three-dimensional structure of soybean trypsin/chymotrypsin Bowman-Birk inhibitor in solution. Biochemistry 31:999–1010
30. Cierpicki T, Otlewski J (2000) Determination of a high precision structure of a novel protein, Linum usitatissimum trypsin inhibitor (LUTI), using computer-aided assignment of NOESY cross-peaks. J Mol Biol 302:1179–1192
31. Hyberts SG, Goldberg MS, Havel TF, Wagner G (1992) The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci 1:736–751
32. Christendat D et al (2000) Structural proteomics of an archaeon. Nat Struct Biol 7:903–909
33. Kohda D, Hatanaka H, Odaka M, Mandiyan V, Ullrich A, Schlessinger J, Inagaki F (1993) Solution structure of the sh3 domain of phospholipase c-gamma. Cell 72:953–960
34. Sönnichsen FD, DeLuca CI, Davies PL, Sykes BD (1996) Refined solution structure of type III antifreeze protein:hydrophobic groups may be involved in the energetics of the protein-ice interaction. Structure 4:1325–1337
35. Schaffer ML, Deshayes K, Nakamura G, Sidhu S, Skelton NJ (2003) Complex with a phage display-derived peptide provides insight into the function of insulin-like growth factor I. Biochemistry 42:9324–9334
36. Yee A et al (2002) An NMR approach to structural proteomics. Proc Natl Acad Sci USA, 99:1825–1830
37. Pineda-Lucena A, Liao J, Wu B, Yee A, Cort JR, Kennedy MA, Edwards AM, Arrowsmith CH (2002) NMR structure of the hypothetical protein encoded by the YjbJ gene from Escherichia coli. Proteins 47:572–574
38. Abajian C, Yatsunyk LA, Ramirez BE, Rosenzweig AC (2004) Yeast cox17 solution structure and copper(I) binding. J Biol Chem 279:53584–53592
39. Arnesano F, Balatri E, Banci L, Bertini I, Winge DR (2005) Folding studies of cox17 reveal an important interplay of cysteine oxidation and copper binding. Structure 13:713–722

40. Cooke RM, Harvey TS, Campbell ID (1991) Solution structure of human insulin-like growth factor 1: a nuclear magnetic resonance and restrained molecular dynamics study. Biochemistry 30:5484–5491
41. Dalgarno DC, Botfield MC, Rickles RJ (1997) SH3 domains and drug design:ligands, structure, biological function. Biopolymers 43:383–400
42. Huang X, Miller W (1991) A time-efficient, linear-space local similarity algorithm. Adv Appl Math 12:337–357
43. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448
44. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X, version 2.0. Bioinformatics 23:2947–2948
45. Holm L, Park J (2000) DaliLite workbench for protein structure comparison. Bioinformatics 16:566-567
46. Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104:59–107
47. Kyte J and Doolittle R F (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132
48. Eisenberg D, Weiss R M, Terwilliger T C, Wilcox W (1982) Hydrophobic moments and protein structure. Faraday Symp Chem Soc 17:109–120
49. Engelman DM, Zaccai G (1986) Bacteriorhodopsin is an inside-out protein. Proc Natl Acad Sci USA 77:5894–5898
50. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78:3824–3828
51. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838
52. Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3:842–848
53. Wolfender R, Anderson L, Cullis PM, Soulhgate CC (1981) Affinities of amino acids side chains for solvent water. Biochemistry 20:846–855
54. http://www.statsoft.com/?kw=spss&gclid=ck3ik-namzscfcitzaodhfy-ig
55. Spearman C (1906) General intelligence, objectively determined and measured. Am J Psychol 6:201–293

56. Vajdos FF, Ultsch M, Schaffer ML, Deshayes KD, Liu J, Skelton NJ, de Vos AM (2001) Crystal structure of human insulin-like growth factor-1:detergent binding inhibits binding protein interactions. Biochemistry 40:11022–11029
57. Koepke J, Ermler U, Warkentin E, Wenzl G, Flecker P (2000) Crystal structure of cancer chemopreventive bowman-birk inhibitor in ternary complex with bovine trypsin at 2.3 a resolution. Structural basis of janus-faced serine protease inhibitor specificity. J Mol Biol 298:477–491
58. Raymond JA, DeVries AL (1972) Freezing behavior of fish blood glycoproteins with antifreeze properties. Cryobiology 9:541–547
59. Raymond JA, DeVries AL (1977) Adsorption inhibition as a mechanism of freezing resistance in polar fishes. Proc Natl Acad Sci USA 74:2589–2593
60. Li Q, Luo L (1993) The kinetic theory of thermal hysteresis of a macromolecule solution. Chem Phys Lett 216:453–457
61. Li Q, Luo L (1994) Further discussion on the thermal hysteresis of the ice growth inhibitor. Chem Phys Lett 223:181–184
62. Hall DG and Lips A (1999) Phenomenology and mechanism of antifreeze peptide activity. Langmuir 15:1905–1912
63. Kristiansen E, Zachariassen KE (2005) The mechanism by which fish antifreeze proteins cause thermal hysteresis. Cryobiology 51:262–28
64. Chao H, Sönnichsen FD, DeLuca CI, Sykes BD, Davies PL (1994) Structure-function relationship in the globular type iii antifreeze protein: identification of a cluster of surface residues required for binding to ice. Protein Sci 3:1760–1769
65. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. Adv Protein Chem 14:1–63
66. Minervini G, Evangelista G, Polticelli F, Piwowar M, Kochanczyk M, Flis L, Malawski M, Szepieniec T, Wiśniowski Z, Matczyńska E, Prymula K, Roterman I (2008) Never born proteins as a test case for ab initio protein structures prediction. Bioinformation 3:177–179

Tables

Table 1 A list of proteins selected for analysis, divided into six groups: growth factors, serine protease inhibitors, antifreeze proteins, chaperones, proteins of unknown function and SH3 domains (the PDB IDs in parentheses show deposited structures of identical amino acid sequence excluded from the discussion)

Group	Protein name	PDB ID
Growth factors	Insulin-like growth factor 1	1BQT, (1PMX, 2GF1, 3GF1)
Serine protease inhibitors	Trypsin/chymotripsin Bowman-Birk inhibitor	1BBI, (2BBI)
	Eglin C	1DWM, 1EGL
Antifreeze proteins	Antifreeze protein type III	1KDE, 1MSI
Proteins of unknown function	Protein MTH_1184	1GH9
	Putative uncharacterized protein	1RYJ
	UPF0337 protein yjbJ	1RYK
	Putative uncharacterized protein	1YVC
	HEF-like protein	2CRE
	Zinc finger CW-type PWWP domain protein 1	2E61
	UPF0165 protein AF_2212	2NWT
Chaperones	Cytochrome c oxidase copper chaperone	1U97, (1U96, 1Z2G)
	Mitochondrial import inner membrane translocase subunit TIM14	2GUZ
SH3 domains	Phospholipase C-gamma	1HSQ
	Myosin-3 isoform	1RUW (1VA7)
	Obscurin	1V1C
	CRK-associated substrate	1WYX
	SH3-domain kinase binding protein 1	2DA9
	Pro-Ser-Thr phosphatase-interacting protein	2DIL
	SH3 multiple domains 1	2DNU

Table 2 SE [bit] and *I* [bit] characteristics of proteins representing growth factors, serine protease inhibitors, antifreeze proteins, chaperones, proteins of unknown function, proteins similar to 2CRE and SH3 domains.

Group	PDB ID:chain	SE_+	SE_+^{max}	$SE_+^{max} - SE_+$	I_+	SE_-	SE_-^{max}	$SE_-^{max} - SE_-$	I_-
Growth factors	1BQT:A	2.24	3.17	0.92	18.99	2.81	3.32	0.51	32.82
Serine protease inhibitors	1BBI:A	2.70	3.46	0.76	40.55	1.88	3.32	1.44	33.69
	1DWM:A	2.91	3.17	0.26	31.71	2.62	3.00	0.38	28.29
	1EGL:A	2.53	3.00	0.47	22.75	2.48	3.00	0.52	22.75
Antifreeze proteins	1KDE:A	2.76	3.17	0.41	32.85	2.26	3.17	0.91	37.64
	1MSI:A	2.82	3.32	0.50	38.80	2.59	3.46	0.87	35.63
Chaperones	1U97:A	1.87	3.00	1.13	15.40	2.49	3.17	0.68	34.06
	2GUZ:A	2.84	3.32	0.47	31.92	2.47	3.32	0.85	35.24
Proteins of unknown function	1GH9:A	2.05	3.32	1.27	38.58	2.30	3.42	1.02	39.40
	1RYJ:A	2.82	3.46	0.64	38.41	3.08	3.58	0.50	43.55
	1RYK:A	2.98	3.46	0.47	43.49	2.99	3.46	0.47	31.07
	1YVC:A	2.28	3.32	1.04	24.10	3.11	3.46	0.35	42.27
	2E61:A	2.42	3.00	0.57	29.93	2.47	3.17	0.70	29.52
	2NWT:A	0.75	2.58	1.83	8.94	1.84	2.81	0.97	21.72
	2CRE:A	3.25	4.00	0.75	57.23	3.40	4.09	0.68	54.92
Proteins similar to 2CRE	2DA9:A	3.15	3.33	0.17	35.34	2.88	3.46	0.57	45.32
	1U5S:A	3.36	4.00	0.64	61.50	3.43	4.00	0.57	54.44
	1U5S:B	1.96	3.00	1.04	26.71	2.29	3.17	0.87	36.41
SH3 domains	1HSQ:A	3.17	3.59	0.42	36.28	2.70	3.59	0.88	41.98
	1RUW:A	2.37	3.00	0.63	30.32	2.78	3.00	0.22	25.80
	1V1C:A	2.99	3.17	0.19	30.61	2.75	3.17	0.43	32.87
	1WYX:A	2.68	3.33	0.66	27.70	2.80	3.32	0.52	38.56
	2DA9:A	3.15	3.32	0.17	35.35	2.89	3.46	0.57	45.33
	2DIL:A	2.82	3.58	0.77	34.29	3.11	3.70	0.59	37.50
	2DNU:A	3.02	3.70	0.68	38.05	3.25	3.81	0.57	55.27

Table 3 Sequence similarity for two pairs of proteins. Two proteins (2DA9, 1U5S) with known biological activity are compared to the one (2CRE) of unknown biological function

	2CRE – 2DA9	2CRE – 1U5SA
Identity	42.3 %	31.9 %
Similarity	77.5 %	75.4 %
Number of gaps	1	3
Sequence alignment score (Waterman-Eggert)	139	121
RMS-D [Å]	3.5	2.2

Table 4 Correlation coefficients for $\Delta\tilde{H}_j$ profiles of proteins from the group of SH3 domains. The significant correlations are given in bold. N – number of points, R – Spearman's rank correlation coefficient, p – p value

	1RUW	**1WYX**	**1V1C**	**1HSQ**	**2DIL**	**2DNU**
1WYX	N=68 R=0.1085 p=0.3784					
1V1C	N=68 R=0.0360 p=0.7708	N=68 R=0.1137 p=0.3558				
1HSQ	**N=68** **R=0.2545** **p=0.0362**	N=68 R=0.0668 p=0.5883	N=68 R=-0.0654 p=0.5960			
2DIL	N=69 R=0.1204 p=0.3243	N=68 R=0.1703 p=0.1650	N=68 R=-0.2379 p=0.0508	N=69 R=0.0118 p=0.9236		
2DMU	N=68 R=0.0016 p=0.9897	N=68 R=0.0442 p=0.7202	N=68 R=-0.0747 p=0.5449	N=70 R=0.1323 p=0.2749	**N=69** **R=0.5810** **p=0.0000**	
2DA9	N=68 R=-0.0021 p=0.9867	N=68 R=0.1974 p=0.1066	N=68 R=0.1362 p=0.2682	N=70 R=0.1731 p=0.1519	**N=69** **R=0.2861** **p=0.0172**	**N=70** **R=0.3910** **p=0.0008**

Table 5 Correlation coefficients measuring relationship between similarities of sequences and 3-D structures, and $\Delta\tilde{H}_j$ profiles. R – Spearman's rank correlation coefficient, p – p value. The significant values are given in bold

	Waterman-Eggert score	$\lvert\Delta SE_+\rvert$	$\lvert\Delta SE_+^{\max}\rvert$	$\lvert\Delta(SE_+^{\max} - SE_+)\rvert$	$\lvert\Delta I_+\rvert$	$\lvert\Delta SE_-\rvert$	$\lvert\Delta SE_-^{\max}\rvert$	$\lvert\Delta(SE_-^{\max} - SE_-)\rvert$	$\lvert\Delta I_-\rvert$
$\lvert\Delta SE_+\rvert$	**R=-0.1490 p=0.0459**								
$\lvert\Delta SE_+^{\max}\rvert$	R=0.0789 p=0.2925	**R=0.4653 p=0.0000**							
$\lvert\Delta(SE_+^{\max} - SE_+)\rvert$	R=-0.1456 p=0.0512	**R=0.5635 p=0.0000**	R=0.0772 p=0.2012						
$\lvert\Delta I_+\rvert$	R=0.0261 p=0.7284	**R=0.4936 p=0.0000**	**R=0.6099 p=0.0000**	**R=0.1762 p=0.0033**					
$\lvert\Delta SE_-\rvert$	**R=-0.1979 p=0.0077**	**R=0.2415 p=0.0001**	**R=0.4700 p=0.0000**	**R=0.1423 p=0.0180**	**R=0.3055 p=0.0000**				
$\lvert\Delta SE_-^{\max}\rvert$	R=0.0929 p=0.2150	**R=0.2540 p=0.0000**	**R=0.8295 p=0.0000**	R=-0.0084 p=0.8890	**R=0.4743 p=0.0000**	**R=0.4688 p=0.0000**			
$\lvert\Delta(SE_-^{\max} - SE_-)\rvert$	**R=-0.2318 p=0.0017**	R=0.0356 p=0.5556	R=-0.0850 p=0.1592	R=0.0938 p=0.1200	R=-0.0932 p=0.1223	**R=0.3613 p=0.0000**	R=-0.0945 p=0.1173		
$\lvert\Delta I_-\rvert$	R=0.1101 p=0.1413	**R=0.2131 p=0.0004**	**R=0.5738 p=0.0000**	R=-0.0331 p=0.5843	**R=0.3413 p=0.0000**	**R=0.3520 p=0.0000**	**R=0.7045 p=0.0000**	R=-0.0810 p=0.1796	
RMS-D [Å]	**R=-0.2798 p=0.0467**	R=0.1903 p=0.1452	**R=0.2635 p=0.0419**	R=0.1005 p=0.4450	R=0.2370 p=0.0683	**R=0.4217 p=0.0008**	R=0.1302 p=0.3214	**R=0.2937 p=0.0228**	R=0.2065 p=0.1135

Figure captions

Fig. 1 $\Delta\tilde{H}_j$ profile of 1BQT representing the group of growth factors with residues responsible for biological function (interaction with the IGF-1 antagonist F1 as observed in 1PMX) presented as green points (**a**), its surface representation (**b**) and ribbon model with residues engaged in biological function presented as spheres (**c**). The color scale on the right is applied to differentiate the $\Delta\tilde{H}_j$ values in 3-D representations. The values of correlation coefficients calculated pair-wise for all proteins belonging to this group of proteins are given to show the strong similarity between the $\Delta\tilde{H}_j$ profiles

Fig. 2 $\Delta\tilde{H}_j$ profiles of proteins representing the group of serine protease inhibitors with residues (42-48) of eglin C (1DWM, 1EGL) that take part in protein-protein interactions (observed in 1CSE, 1SIB or 1TEC) presented as green points. The correlation coefficient (and p value to estimate the significance) for these two profiles is given as well (**a**) Ribbon models of these proteins with residues of eglin C (1DWM, 1EGL) that take part in protein-protein interactions (observed in 1CSE, 1SIB or 1TEC) presented as spheres (**b**). The color scale shown on the right of the graph is applied to 3-D representations

Fig. 3 $\Delta\tilde{H}_j$ profiles of proteins representing antifreeze proteins with key residues interacting with ice. The correlation coefficient (and p value to estimate the significance) for these two profiles is given as well (**a**). Surface representations of these proteins (**b**) and ribbon models with key residues interacting with ice, presented as spheres (**c**). The color scale shown on the right of the graph is applied to 3-D representations

Fig. 4 $\Delta\tilde{H}_j$ profiles of proteins representing chaperones with residues forming copper-binding sites for cytochrome c oxidase copper chaperone, 1U97 (observed in 1U96). The correlation coefficient (and p value to estimate the significance) for presented profiles is given as well (**a**). Surface representations of these proteins (**b**) and ribbon models with residues forming copper-binding sites in cytochrome c

oxidase copper chaperone, 1U97 (observed in 1U96) (**c**). The color scale shown on the right of the graph is applied to 3-D representations

Fig. 5 $\Delta\tilde{H}_j$ profiles of protein of unknown biological function (2CRE) aligned on $\Delta\tilde{H}_j$ profiles of SH3 domains of Ruk (2DA9) and NCK adaptor protein 2 (1U5S:A) with vertical green lines denoting gaps in alignment. The correlation coefficients (together with p values to measure the significance) for two variants of alignment of profiles for 2CRE and 2DA9 are given as well. (**a**). Aligned ribbon models of these three structures with residues found as insertions in the alignment shown as ball-and-stick, and surface representations of these structures (**b**). The color scale shown on the right of the graph is applied to 3-D representations

Fig. 6 $\Delta\tilde{H}_j$ profiles of protein representing SH3 domains taking into account the multiple sequence alignment (**a**), standard deviations (SDs) reflecting the dispersion of $\Delta\tilde{H}_j$ values at each position of the alignment (**b**), aligned ribbon models of these structures colored according to $\Delta\tilde{H}_j$ values (**c**) and the ribbon model of the SH3 domain from *S. cerevisiae* Myo3 (1RUW) colored according to SD values (**d**)

Fig.1

Click here to download high resolution image

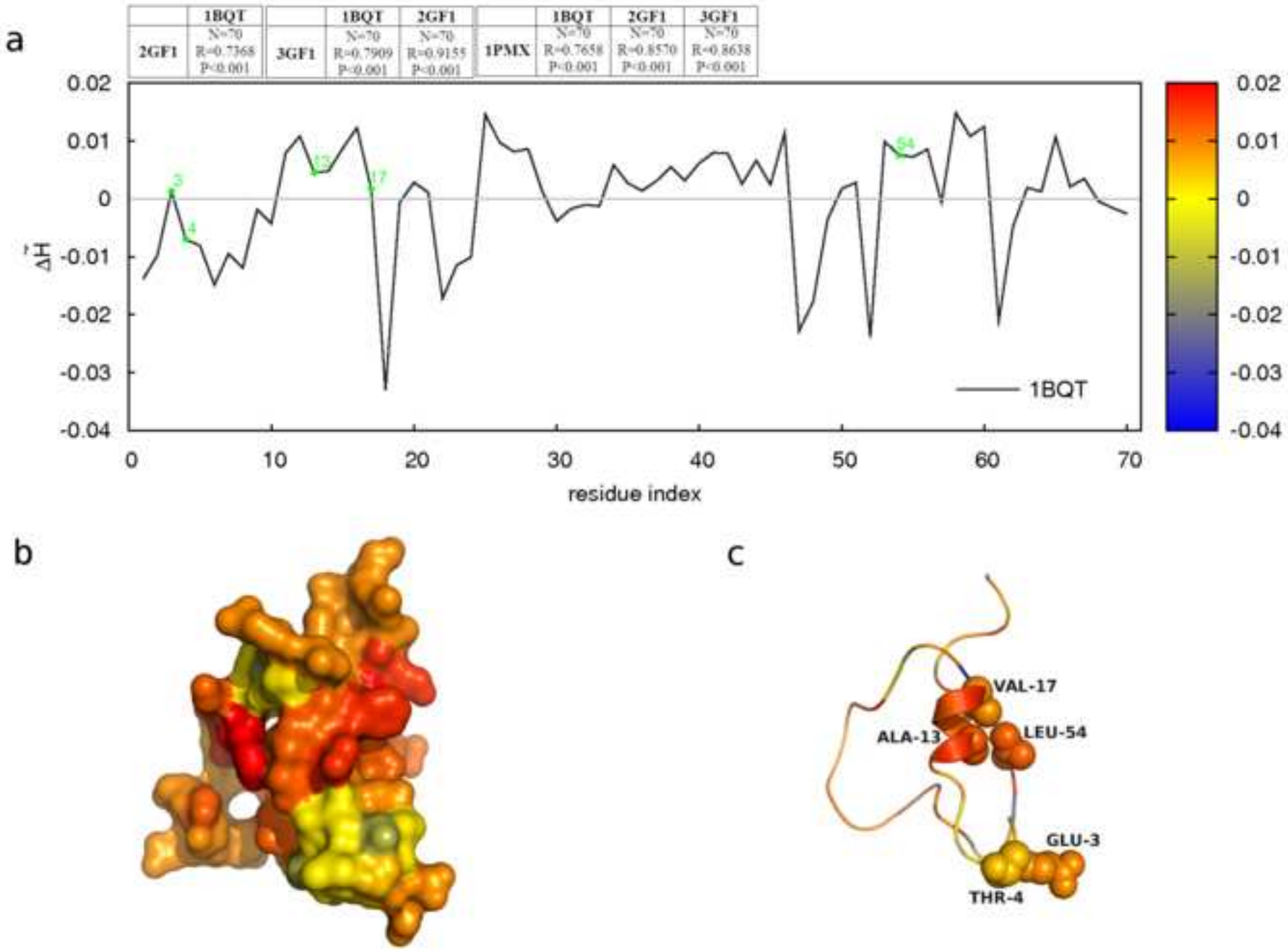

Fig.2

Click here to download high resolution image

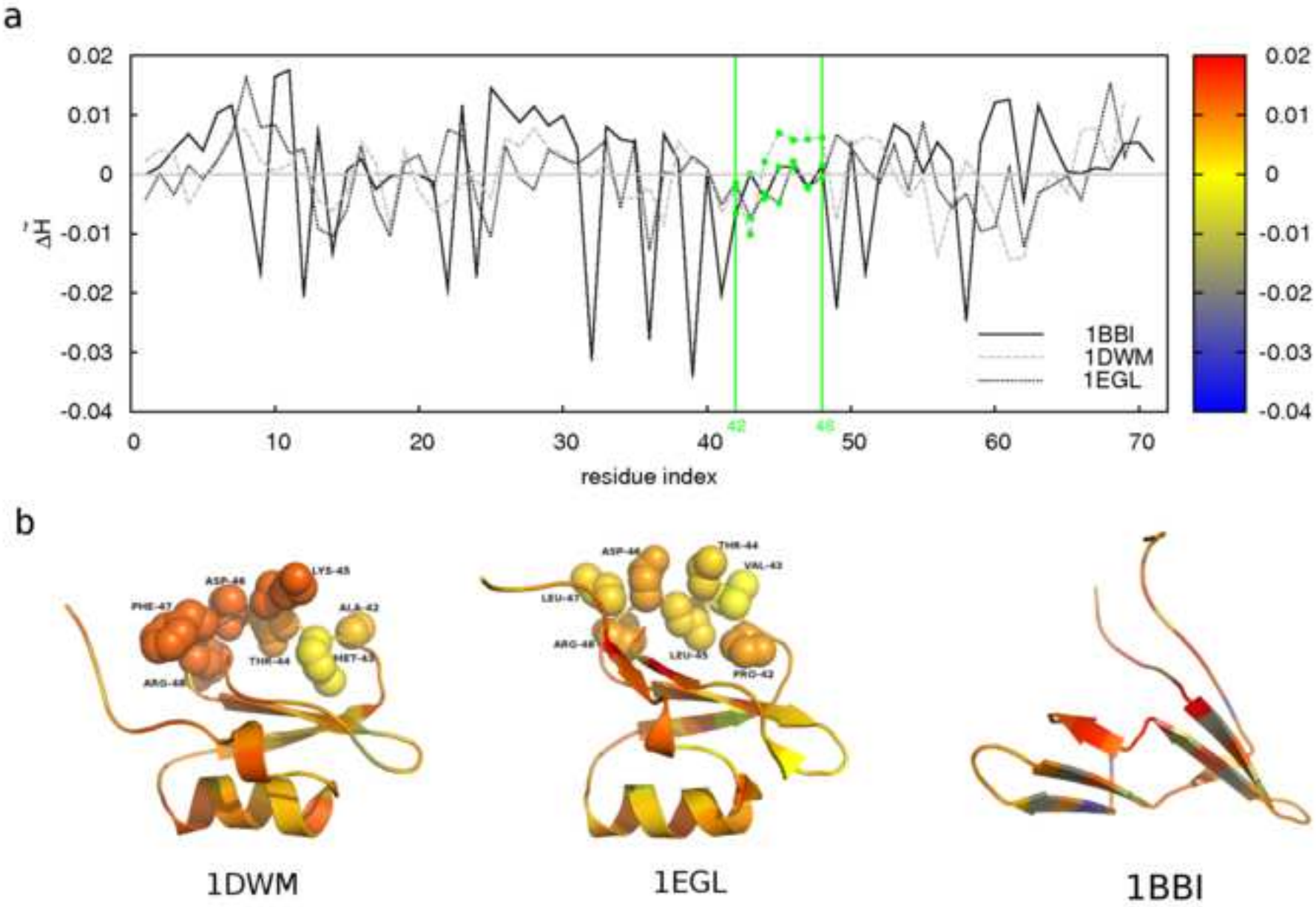

Fig. 3

Click here to download high resolution image

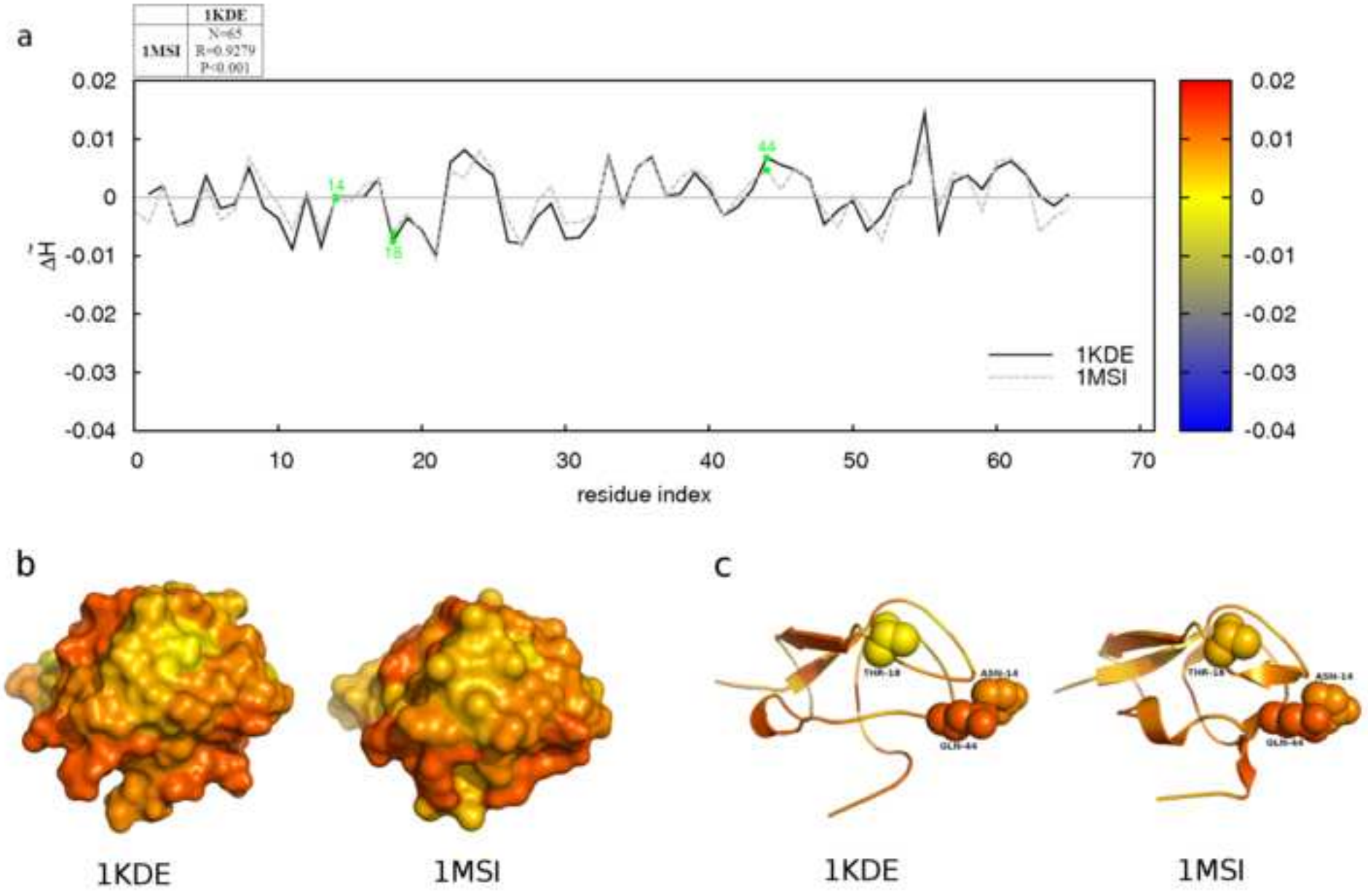

Fig. 4

Click here to download high resolution image

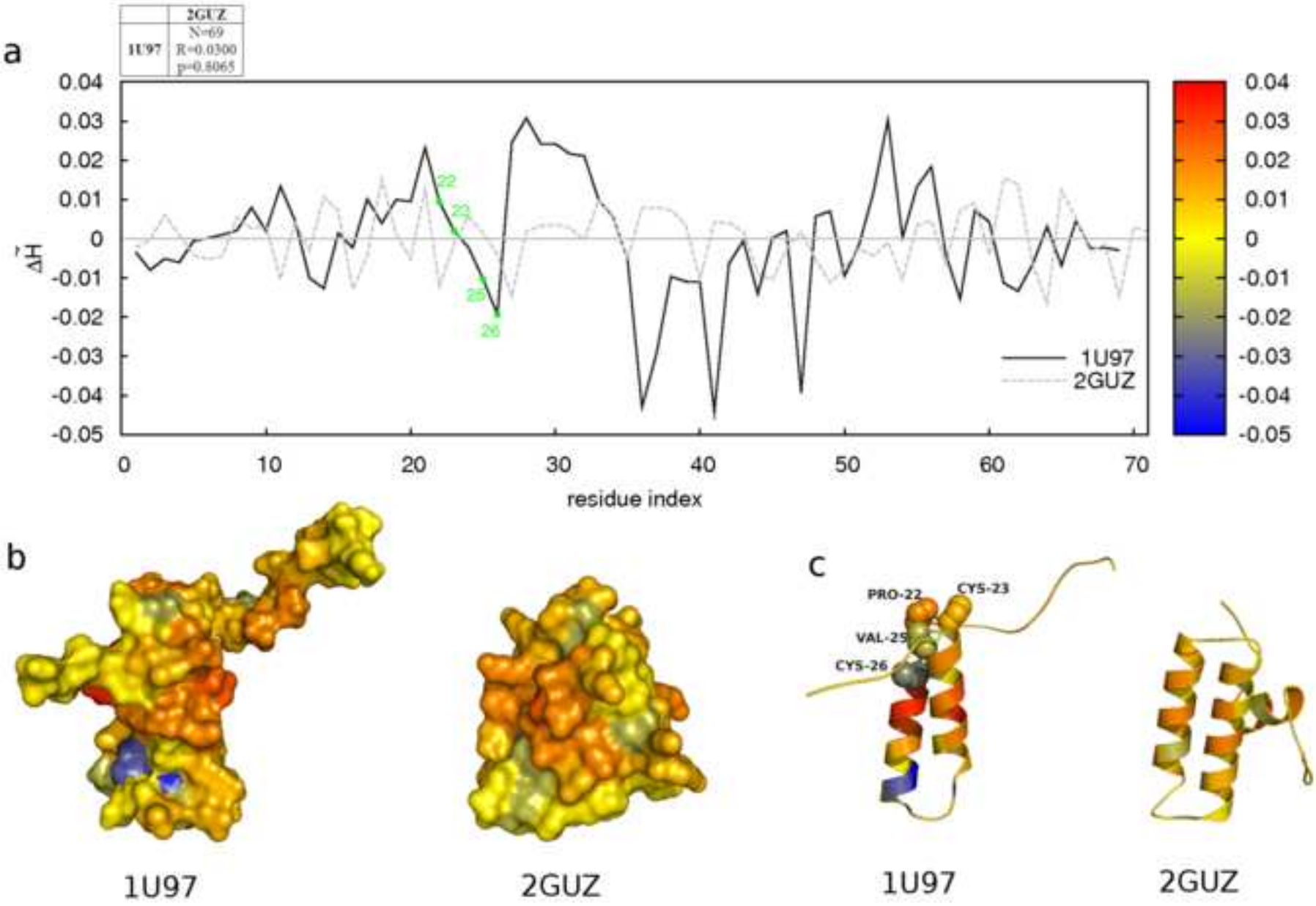

Fig. 5

Click here to download high resolution image

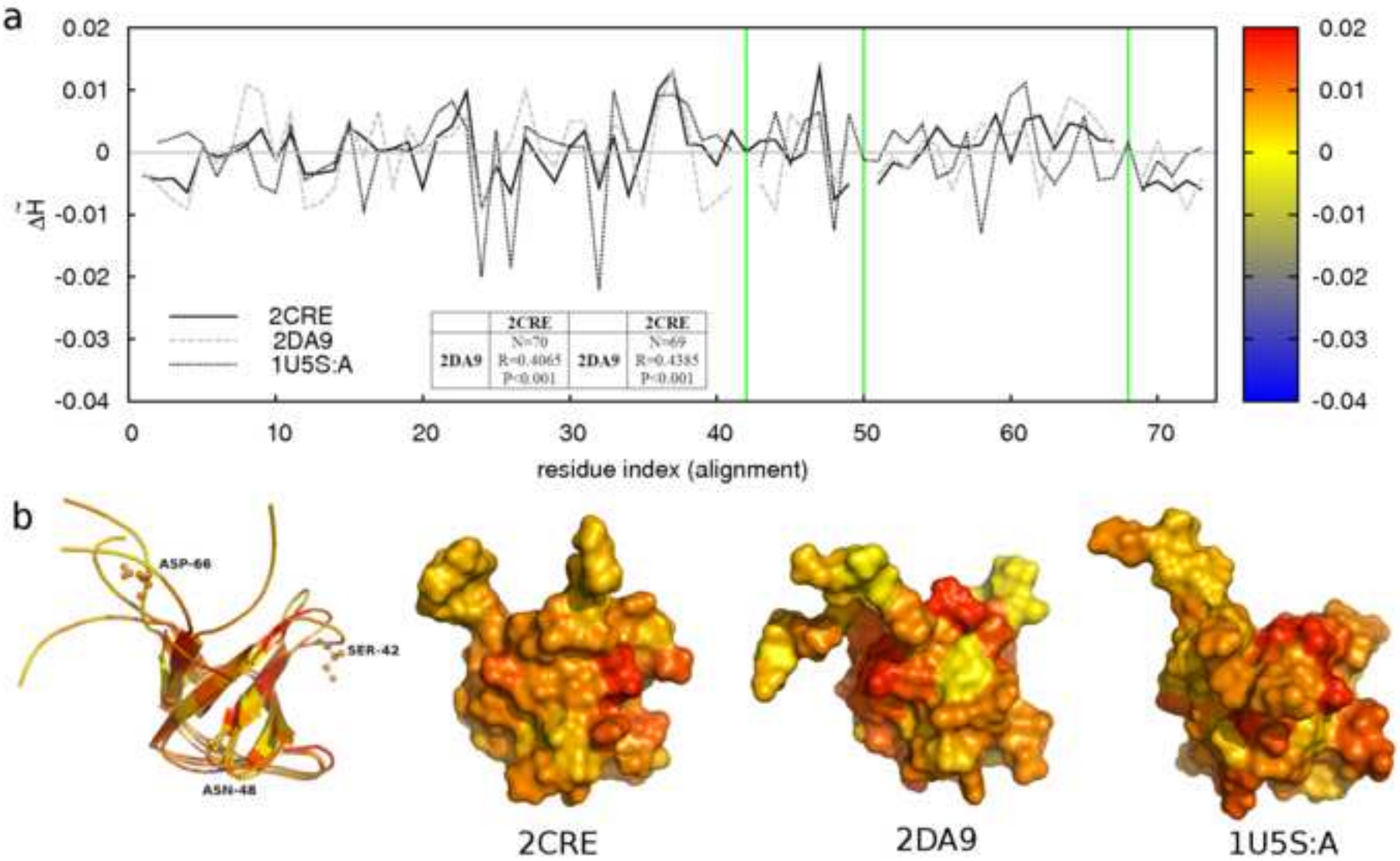

Fig. 6

Click here to download high resolution image