Two-intermediate model to characterise the structure of fast-folding proteins

I. Roterman, L. Konieczny, Wiktor Jurkowski, K. Prymula, M. Banach

www.elsevier.com/locate/yjtbi

Two-intermediate model to characterise the structure of fast-folding proteins

I. Roterman, L. Konieczny, W. Jurkowski, K. Prymula, M. Banach

PII: S0022-5193(11)00269-4
DOI: doi:10.1016/j.jtbi.2011.05.027
Reference: YJTBI6491

To appear in: *Journal of Theoretical Biology*

Received date: 4 December 2010
Revised date: 17 May 2011
Accepted date: 18 May 2011

Cite this article as: I. Roterman, L. Konieczny, W. Jurkowski, K. Prymula and M. Banach, Two-intermediate model to characterise the structure of fast-folding proteins, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2011.05.027

TWO-INTERMEDIATE MODEL TO CHARACTERISE THE STRUCTURE OF FAST-FOLDING PROTEINS

[1,*]Roterman I, [1,2]Konieczny L, [1]Jurkowski W, [1,3]Prymula K, [1,4]Banach M.

1 Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Lazarza 16, 31-530 Krakow, Poland; 2 Chair of Medical Biochemistry, Jagiellonian University – Medical College, Kopernika 7, 31-034 Krakow, Poland; 3 Faculty of Chemistry – Jagiellonian University, Ingardena 3 30-060 Krakow, Poland; 4 Faculty of Physics, Astronomy and Applied Computer Science – Jagiellonian University, Reymonta 4, 30-059 Krakow, Poland

[*] Corresponding Author

Abstract:

This paper introduces a new model which enables researchers to conduct protein folding simulations. A two-step *in silico* process is used in the course of structural analysis of a set of fast-folding proteins. The model assumes an early stage (ES) that depends solely on the backbone conformation, as described by its geometrical properties – specifically, by the V-angle between two sequential peptide bond planes (which determines the radius of curvature, also called R-radius, according to a 2nd degree polynomial form). The agreement between the structure under consideration and the assumed model is measured in terms of the magnitude of dispersion of both parameters with respect to idealized values. The second step, called late-stage folding (LS), is based on the "fuzzy oil drop" model, which involves an external hydrophobic force field described by a three-dimensional Gauss function. The degree of

conformance between the structure under consideration and its idealized model is expressed quantitatively by means of the Kullback-Leibler entropy, which is a measure of disparity between the observed and expected hydrophobicity distributions. A set of proteins, representative of the fast-folding group – specifically, cold shock proteins – is shown to agree with the proposed model.

Keywords: protein folding, hydrophobicity, information theory

INTRODUCTION

Proteins composed of fewer than 150 amino acids often fold very quickly, i.e. in tens of microseconds (Pande et al. 1998; Englander 2000). The presence of hidden intermediates in this process has been postulated on the basis of experimental observations, as predicted by the funnel model and extensive folding process kinetics studies (Ozkan et al 2002; Mayor et al 2000; Clarke et al 1994). The super-folding protein GFP model was developed to validate the fast-folding mechanism experimentally and to verify the influence of mutations on biological function (Fisher and DeLisa 2008). The impact of crowding upon the protein folding process, which stresses the critical importance of surrounding polypeptides, was computationally verified by (Jefferys et al 2010). Recently, NMR spin relaxation dispersion experiments have been performed to quantify the mutational effects on kinetics and overcome the limitations of traditional stopped-flow experiments as applied to fast-folding protein kinetics (Cho et al. 2010). A solid background and broad overview of protein folding processes, based on classical phenomena related to the distribution of energy, can be found in (Chou and Scheraga 1982, Chou and Carlacci 1991, Chou et al 1983a, Chou et al. 1983b, Chou et al. 1984, Chou et al. 1990, Chou et al 1992, Chou et al. 1988, Chou et al. 1986). The statistical point of view is introduced in (Chou and Zhang 1993, Chou 1995a, Chou 1995b, Chou and Zhang 1994, Chou and Zhang 1995, Mao et al 1994, Zhang and Chou 1992) while the kinetic point of view

is detailed in (Chou 1990, Chou 1993, Chou and Shen 2009, Shen et al 2009). Molecular dynamics and folding disorder are further discussed in (Wallace 2010, Wang and Chou 2009). In order to perform comparative analysis, we intend to focus on the set of cold shock proteins, as well as some additional proteins (prefoldins and chaperonins).

METHODS

Data – The proteins presented in this paper were selected on the basis of source descriptions. All such proteins are classified as fast-folding (Tab. 1).

A set of ultra-fast folding proteins was derived, based on annotations found in literature (Dyer 2007; Kubelka et al 2004; Ghosh et al. 2007). Representative structures presented in Tab. 1 relate specifically to ultra-fast folding fragments and have been selected on the basis of references (Bogatyreva et al 2009) and the quality of available structural data (Berman et al 2002). The molecular function of selected structures (Tab. 1) refers to entire proteins and is expressed by means of the Gene Ontology published by the UniProt Consortium (2009).

Additionally, a set of cold shock proteins was selected to verify the applicability of the model to proteins with similar biological functions. Such common properties include binding to nucleic acids. Two chaperonins were also included in the study group, as suggested by prior research (Prymula et al 2009) where they were found to exhibit high structural accordance with the presented model.

The properties of proteins selected for analysis with respect to sequential and structural similarities enabled us to define a non-redundant subset (Tab. 2), which was further analyzed with the use of ClustalW. Pair-wise sequence identity below 20% (as reported by ClustalW) was taken as a criterion of alignment (Chenna et al. 2003). In some cases, structures exhibiting higher sequential identity revealed significant differences (RMS-D) and were also included in the study group (Tab. 3) – this is why similar proteins appear in our analysis.

The sequential similarity of the studied proteins is presented in Tab.1S (Supplementary Materials). High sequential similarity does not preclude a protein from being analyzed with the presented model, as even a single mutation may influence the structure of the protein's hydrophobic core (as reported in (Banach et al 2011)).

The model is assumed to be applicable to proteins with polypeptide residue lengths of approximately 150, although longer proteins (such as 3BDN, with 236 amino acids) were also analyzed to estimate the applicability of the model to proteins whose size exceeds the initial assumptions (Stayrook et al 2008).

Short polypeptide chains (1RIJ_A) represent domains which may be treated as independent units when analyzing the folding process.

Two-step protein folding process

The protein folding process has been shown experimentally to involve multiple steps, along with an unknown number of intermediates (Roterman 2007; Jurkowski et al 2004). The model presented in this work assumes a two-step process:

U => ES => LS => N

Where: U – unfolded, ES – early stage, LS – late stage and N – native structural form.

Early stage model. This model assumes the dominant role of the backbone whose conformation is expressed by two geometric parameters (Roterman 1995). The first one is called V-angle – the dihedral angle between two sequential peptide bond planes, the value of which is close to 0º for helical forms and close to 180º for extended and β-like structures. The second parameter, which appears to be determined to some extent by the first one, is the radius of curvature of the polypeptide fragment (pentapeptide), which is small for helical structures (according to the applied model, the radius of curvature for helical structures is

approximately 2Å) and large for β-structural forms (in the case of linear forms, the radius of curvature is theoretically infinite; according to calculations presented in this paper the actual radius of curvature for structures with a V-angle of 180º was found to be in the range of 7 to 15 on the $\log_2$ scale). The relation between both parameters, which may be expressed using a second-degree polynomial

$$\ln(R) = 0.00034V^2 - 0.02009V + 0.848$$

determines the optimal path on the Ramachandran plot (which represents the complete conformational space). An elliptical path on the Phi-Psi map links the locations of all secondary structures. This path is assumed to represent the limited conformational subspace available to the backbone in the ES step of the folding process. Agreement between the model and the actual protein is estimated in terms of the average distance ($D_{average}$) between the projected and observed curvature radius for a specific V-angle, affecting a particular residue in the polypeptide chain. A visual interpretation of the ES model is presented in Fig. 1.

Late stage model. The tertiary structure of the protein in the LS step of the folding process is assumed to involve a hydrophobic core, along with optimization of all other non-bonding interactions (electrostatic, van der Waals and torsion potential). The presence of an external force field is reflected by a three-dimensional Gauss function (Konieczny et al 2006). This model extends the original concept presented by Kauzmann (Kauzmann 1959). The force field simulates the hydrophobic core postulated by the "fuzzy oil drop" model where the highest concentration of hydrophobicity is observed at the center of the ellipsoid, decreasing along with distance from the center and reaching zero on the surface of the "drop", according to Gauss' formula:

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

where $\bar{x}, \bar{y}, \bar{z}$ are the coordinates of the geometric center of the molecule (usually located at the origin of the coordinate system, where each value is equal to 0). The size of the molecule is expressed by the triplet σ_x, σ_y, σ_z, which is calculated for each molecule individually, provided that the longest possible distance between effective atoms within the molecule coincides with the appropriate coordinate system axis. σ values are defined as 1/3 of the longest distance between two effective atoms along each axis. The value of the Gauss function at any point of the protein body can be treated as the idealized hydrophobicity density, determining the structure of the protein's hydrophobic core.

According to the "fuzzy oil drop" model, idealized hydrophobicity can be calculated at any point with the use of the Gauss function, assuming that the molecule's geometric center coincides with the origin of the coordinate system. In turn, empirical hydrophobicity distribution is given by the function presented by Levitt (Levitt 1976):

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

where N expresses the number of amino acids in the protein (number of grid points), $\tilde{H}_i^r$ expresses the hydrophobicity of the *i-th* residue according to the accepted hydrophobicity scale (the Aboderin scale was applied in this work (Aboderin 1971)), r_{ij} expresses the distance between the *i-th* and *j-th* interacting residues, and c expresses the cutoff distance which, according to the original paper (Roterman 2007), is assumed to be 9Å. The values of $\tilde{Ho}_j$ are

standardized via division by the $\tilde{Ho}_{sum}$ coefficient, which represents the aggregate sum of all hydrophobicity values assigned to grid points.

The distribution of hydrophobicity in the analyzed molecules seems highly consistent with the proposed model. Irregularities observed in certain proteins appear to be target-oriented and related to ligand binding sites or enzymatic active sites.

Kullback-Leibler information entropy. The agreement between the idealized and observed hydrophobicity distribution is measured according to the Kullback-Leibler relative (divergence) entropy (Nalewajski 2006), which quantifies the distance between both distributions. The distance between the observed and the theoretical (O/T) distribution was calculated as part of the presented study. This value can only be analyzed comparatively, with respect to other solutions – thus, random distribution of hydrophobicity (O/R) was also estimated. The relation O/T < O/R was taken as evidence of non-random distribution, closely approximating theoretical values.

$$D_{KL}(p\,|\,p^0) = \sum_{i=1}^{N} p_i \log_2(p_i / p_i^0),$$

where: D_{KL} – distance entropy, p – probability of occurrence of a particular event in the observed distribution (denoted as O – observed distribution in our analysis) p^0 – corresponding probability in the reference distribution (denoted as T – theoretical or R – random distribution in our analysis),. The index i corresponds to a particular amino acid, while N denotes the total number of amino acids in the polypeptide chain.

RESULTS

Selected proteins have been analyzed to assess the applicability of the proposed model. Early-stage conformance is not expected to remain evident in the native form of the protein, as the

LS step may significantly alter its structure, erasing the characteristics of the earlier stage. Despite this phenomenon, the shape of some proteins reveals significant contributions of ES elements, mostly in the scope of well-defined fragments of the secondary structure.

ES model applicability

Mean values expressing distance (versus the approximation function) are given in Tab. 4. Proteins with $D_{average} < 1.0$ were treated as consistent with the model (Fig. 2).

Results for 2HAX and 1CSA are presented here in order to provide two examples which do not agree with the ES model. The distributions of points representing their V-versus-ln(R) characteristics is shown in Fig. 2.

High irregularity of the 2HAX molecule, as compared with the model, can be explained by its significant involvement in interactions with ligands, RNA molecules and other proteins. The ratio of noninteracting-versus-interacting residues is relatively low. External molecules interacting with the 2HAX chain further distort the original structure of this protein, exacerbating the observed irregularities. The presence of ES elements in the final native structure, while unexpected, cannot be ruled out. The 1LMB protein (also involved in DNA complexation) crystallizes in dimeric form while retaining structural agreement with both ES and LS models.

The 1CSQ protein, representative of the β-barrel structure, was arbitrarily selected to present another example of a protein which does not satisfy the ES model assumptions.

The results of applying the ES step are shown on the examples of 1IET and 2ZDI, both of which represent good agreement with the assumed model (Fig. 3). In 2ZDI only seven residues diverge significantly from expectations, due to the high percentage of helical structures. Some residues accordant with expectations are involved in protein-protein interactions which do not, however, affect the structure of the protein (thus, the relevant residues remain in good agreement with the ES model). The 1IET protein, whose polypeptide

chain length is similar to that of 2ZDI, represents a wider distribution of points, exhibiting high structural homogeneity with 2ZDI.

The $D_{average}$ values for both proteins appears to be lower than 1 unit. The structure of 1IET, although quite diverse (involving helical, β-structural and random coil forms), seems to agree with the ES model.

A 3D representation of selected protein structures, demonstrating the applicability of the ES model, is shown in Fig. 4. (2HAX and 1CSQ) and Fig 5. (2ZDI and 1IET). Residues discordant with the ES model are highlighted in order to show their placement in the native structural form of the protein.

The structure of 2ZDI includes a bihelical twist and therefore approximates an elongated fibrillar form, quite different from globular drop-like proteins. This is why the molecule, while in agreement with the ES model, diverges from the LS model.

The 1IET protein, which has a mixed secondary structure, appears to be accordant with the ES model (on the basis of the $D_{average}<1$ criterion).

LS model applicability

The distribution of hydrophobicity density along the polypeptide chains in proteins exhibiting poor structural agreement with the LS model is shown in Fig. 6. (2HAX) and in Fig. 7. (2ZDI). Two proteins – 2HAX and 2ZDI – have been selected to represent cases of poor agreement with the LS model. Both figures contrast the expected distribution of hydrophobicity along the polypeptide chain with empirical observations. Random distribution is also plotted to demonstrate the baseline case.

3D representations of these two proteins can be seen in Fig.8.

Two examples of proteins representing structures accordant with the LS model have been selected: 1IET and 1CSQ. Their hydrophobicity profiles are shown in Fig. 9. and Fig. 10. respectively.

Proteins with a highly differentiated secondary structure (1IET) and a typical β-barrel structure (1CSQ) appear to produce hydrophobic cores accordant with theoretical predictions. 3D representations of 1IET and 1CSQ (proteins exhibiting good agreement with the LS model) are shown in Fig. 11. Differences between the observed and idealized hydrophobicity distributions are relatively narrow and restricted to the center of the 1IET molecule, suggesting that the core's hydrophobicity is somewhat higher than expected (see the profile in Fig. 9).

Accordance of observations with the model

The properties of entire study group are summarized in Tab. 5.

Classification of proteins with respect to the selected criteria is as follows:

1. The presence of ES structural elements in crystalline forms is generally not expected, although certain proteins do exhibit a well-defined secondary structure (mostly helical). In such cases, the presence of an LS structure is not expected. According to the two-step model, proteins representing good agreement with the ES model are assumed to cease folding upon completion of the first step.
2. The presence of LS structural elements is expected on the assumption that the environment (represented by the external force field in the "fuzzy oil drop" model) influences – or indeed determines – the folding process, resulting in formation of a hydrophobic core. All molecules belonging to the group of downhill (fast-folding) proteins support this assumption.

3. Proteins which do not conform to either model (ES or LS) are assumed to be influenced by factors other than backbone conformation and/or immersion in a simple aqueous environment. Typically, they are found in complexes with other protein molecules or ligands (see the rightmost column in Tab. 5).

Most of the proteins in the presented set exhibit good agreement with both models. All proteins recognized as "fast folding" contain a well-defined hydrophobic core. The disappearance of ES intermediates in the LS form is to be expected since LS involves packing of the protein body and therefore alters the conformation of its backbone.

Proteins which appear to diverge from the "fuzzy oil drop" model are highly specific in shape. Two of them (2ZDI and 2ZQM) are entirely helical, which explains their agreement with the ES model and lack of agreement with the "fuzzy oil drop" model. It should be noted that their actual structure is far removed from the idealized "globule". On the other hand, LS discordance exhibited by 3BDN is due to its very large size (although significant involvement of helical elements causes it to conform to the ES model).

The 1HZC protein includes three mutations (versus the WT protein), introduced intentionally to influence its stability. Hence, this artificially constructed protein differs from the standard structural model and cannot be described by either ES or LS. The 2HAX protein, available in PDB, assumes the form of a dimer interacting with a ligand. This phenomenon affects the structure of the protein's polypeptide chain. Such task-oriented irregularities in hydrophobicity distribution are observed in some proteins, enabling highly selective complexation of specific ligands or protein-protein interactions (Brylinski et al 2007b). The 2HAX protein is a member of this group. Its intense involvement in interactions with other molecules may influence the original structure of the protein, explaining its poor agreement with either model (ES or LS).

CONCLUSIONS

The two-step model was assumed to enable *in silico* simulations of the protein folding process. The group of proteins presented in this analysis appears to correspond to the assumed model. The most interesting observation is that the entire group of fast-folding proteins exhibits good accordance with the LS model. The folding process occurs in an aqueous environment and seems influenced only by water. On the other hand, proteins engaged in interaction with other molecules generally diverge from the model, which can be easily explained by external influences on the folding process. Chaperonins (which are expected to fold spontaneously) appear to disagree with the LS model as their forms are highly helical and fibrillar, without any hydrophobic core. It seems that – at least in their case – the folding process halts at the first intermediate, which is represented by ES.

The irregularity of hydrophobicity distribution may result from external factors other than the influence of water upon the folding process. The identification of ligand-binding loci in certain molecules (Brylinski et al 2007a) suggests that the presence of ligands and their active participation in the folding process may produce highly selective binding cavities (Brylinski et al 2007b; Brylinski et al 2006).

It should be noted that the investigated proteins belong to the so-called "easy" group (as named in the CASP classification (Orengo et al 1999)). "Easy" proteins are often correctly simulated *in silico* by various numerical methods. However, the excellent agreement between the observed crystalline structures and theoretical predictions based upon a well-defined model strongly suggests that our generalized model mimics the actual folding process. This observation is additionally supported by analysis of protein domains in large molecules, which exhibit structures highly consistent with theoretical predictions – particularly with regard to the LS model.

In-depth analysis has also been undertaken to select and identify proteins that follow the "fuzzy oil drop" model and to enumerate selection criteria which, when applied to amino acid sequences, might indicate whether a particular protein belongs to this category. The question "why do some proteins fail to follow the natural model of water-influenced folding?" is the next issue to be considered.

Good structural accordance of trans-membrane proteins with the "fuzzy oil drop" model suggests that the influence of the environment, which is a highly variable parameter, should be taken into consideration to a far greater extent than in the past (Zobnina and Roterman 2009). It appears that protein machines, like the 1AON chaperonin (Banach et al 2009), may also be analysed using the "fuzzy oil drop" model. In such a large complex, proteins which are recognized to fold in accordance with the "fuzzy oil drop" model may contribute to the environment for chains which fold later on (as is the case with haemoglobin (Brylinski et al 2007b)). Antifreeze proteins are structurally accordant with the presented model (Prymula et al 2010). Similarly, chaperonins are good examples of the applicability of the presented model to large-scale structural analysis – in fact, some of them have already been recognized as supporting the reliability of the model (Prymula et al 2009).

The results unequivocally show that ES (geometric model assuming the dominant role of backbone conformation) and LS (involving the presence of an external hydrophobic force field in the form of a "fuzzy oil drop") seem to reflect the conditions under which the protein attains its final structural form. Fast-folding proteins were selected so that other environmental factors (e.g. other participating molecules) which influence the folding process could be disregarded.

Both models (ES and LS) provide fresh insight into the mechanisms of fast protein folding and may furthermore be adapted for *in silico* folding simulations.

Significant progress has recently been made (including the introduction of pseudo-amino acid compositions (Chou 2010, Lin and Ding 2011)) to improve protein folding rate predictions (Guo 2010) and develop graphic rules (Chou 1989, Chou 1990) to investigate protein folding rates (Chou and Shen 2009b, Shen 2009). The publication Chou 2010, summarizing the 50-year history of scientific approaches to the protein structure problem, shall be mentioned and underlined (Chou 2010).

The protein folding problem is of critical importance for structural biochemistry. Many groups are currently involved in research focusing on protein structure prediction. Arguably the best assessment of progress in this field is offered by the CASP project (http://predictioncenter.org/). A summary of extensive research efforts in the area of protein folding may also be found in (Chou and Zhang 1995, Chou and Shen 2008, Zakeri et al. 2011).

The model presented in this paper is intended for application in protein folding simulations. Thus far, its reliability has been verified on proteins whose structure follows the assumed criteria. Fast-folding proteins have been found to possess a well-defined hydrophobic core. The LS model has been applied in test simulations involving haemoglobin (Brylinski et al 2007b) and ribonuclease (Brylinski et al. 2006). Experience with fast-folding proteins encourages the authors to further apply the model to large-scale protein folding simulations.

Cross-validation of the model's applicability will include statistical predictions, particularly using the jackknife cross validation method (Chou, 2011) widely applied by investigators when examining the accuracy of various models and predictors (Chen et al 2009, Ding et al 2009, Kandaswamy et al. 2011, Liu and Jia 2010, Masso and Vaisman 2010, Mohabatkar 2010, Zeng et al 2009). While this study applied an independent test data set in order to reduce computational complexity, the above actions are foreseen in the future.

It should be noted that public, user-friendly Web interfaces facilitate the development of useful models, methods and predictors (Chou and Shen, 2009a). Thus, care has been taken to provide Web-based access to the methods presented in this paper. Web servers implementing specific elements of the model (ES intermediate, recognition of active sites) are already available – see http://bioinformatics.cm-uj.krakow.pl/beta/index.php/Main_Page.

Acknowledgements

The work was financially supported by Jagiellonian University – Medical College – project K/ZDS/001531. We would also like to thank Piotr Nowakowski from ACC CYFRONET AGH for proofreading the document.

Figure captions:

Fig. 1. The ES model definition. a**)** the Ramachandran plot with low-energy area indicated; b) the relation between the V-angle (dihedral angle between two sequential peptide bond planes) and R – radius of curvature (presented using a logarithmic scale to better depict large values for β-structural forms) as calculated for structures belonging to low-energy fragments of the Ramachandran plot (shown in a) together with the approximation function (2^{nd} degree polynomial); c) the Ramachandran plot with points representing structures accordant with the approximation function shown in b); d) the elliptical path assumed to represent the limited conformational subspace for the early-stage (ES) intermediate; e) the elliptical path linking all secondary structures.

Fig. 2. The ES model as applied to two proteins which represent poor agreement with this model. The dark blue line represents the theoretical idealized dependence between V-angle and Ln(R) – radius of curvature. Pink squares show the results for a particular

protein. Yellow triangles denote the residues whose geometric parameters differ by more than 1 unit (Y-axis scale). The points in blue circles indicate residues not engaged in any interaction with external molecules (ligand, RNA/DNA, protein complexation) in 2HAX.

Fig. 3. Two proteins satisfying the predictions of the ES model. Only 7 amino acids in 2ZDI differ significantly (by more than 1 unit) with respect to Ln(R). The residues engaged in protein-protein interaction are marked by blue circles. The dark blue line represents the theoretical dependence between V-angle and Ln(R) – radius of curvature. Pink squares show this dependence as it appears in both proteins.

Fig.4. 3D view of 2HAX (left) and 1CSQ (right). Fragments marked in blue represent residues which do not agree with the ES model. In 2HAX the residues marked in red do not interact with any ligand or protein – this shows that the majority of this protein is involved in interactions with external molecules.

Fig.5. 3D view of 2ZDI (top) and 1IET (bottom). The fragments marked in white represent residues which do not agree with the ES model. In 2ZDI the residues marked in red interact with other chains in the protein complex. The terminal residues of β-structural fragments and some loops in 1IET, while exhibiting significant structural variances (helical, β-structural and random coil elements), appear to disagree with the ES model.

Fig.6. The theoretical (dark blue symbols – T) and observed (pink squares – O) hydrophobicity density distribution in 2HAX. Random distribution is shown by yellow triangles (R) (top). Residues engaged in ligand complexation are additionally marked by light blue dots, while those involved in protein-protein interaction are marked by dark asterisks. Brown rhombuses represent residues engaged in RNA complexation.

The distribution of D_{KL} values in 2HAX is shown in the lower part of the figure. This includes distance entropy values for the observed (O) versus theoretical (T) distribution,

which is treated as a reference (dark blue), and distance entropy for the observed (O) versus random (R) distribution, shown in pink.

Fig.7. The theoretical (T – dark blue symbols), observed (O – pink squares) and random (R – yellow triangles) hydrophobicity distribution in 2ZDI (top). The figure depicts the D_{KL} profile calculated for the observed distribution versus the theoretical one (dark blue symbols) and for the observed distribution versus random distribution (pink squares). This is an example of poor agreement with the LS model.

The D_{KL} distribution (lower part of the figure) expressing the distance between observed (O) and theoretical (T) values (treated as a reference) is shown in dark blue, while the distance between the observed (O) and random (R) values (treated as a reference) is shown in pink.

Fig.8. Two proteins (2ZDI – top and 2HAX – bottom) representing examples of poor agreement with the LS model. White fragments correspond to DKL values greater than 0.02. The residues marked in red (in 2HAX) are not engaged in any complexation with ligands or other proteins – this shows that the majority of residues in this molecule are, in fact, involved in interaction with other molecules.

Fig.9. The theoretical (T), observed (O) and random (R) distribution of hydrophobicity density in 1IET. Dark blue symbols represent the theoretical distribution, pink squares represent the observed distribution while yellow triangles represent random distribution.

The D_{KL} distribution (lower part of the figure) expressing the distance between observed (O) and theoretical (T) values (treated as a reference) is shown in dark blue, while the distance between the observed (O) and random (R) values (treated as a reference) is shown in pink.

Fig.10. The theoretical (T), observed (O) and random (R) distribution of hydrophobicity density in 1CSQ. Dark blue symbols represent the theoretical distribution, pink squares

represent the observed distribution while yellow triangles represent random distribution. The D_{KL} distribution (lower part of the figure) expressing the distance between observed (O) and theoretical (T) values (treated as a reference) is shown in dark blue, while the distance between the observed (O) and random (R) values (treated as a reference) is shown in pink.

Fig.11. 3D view of two proteins: 1IET (top) and 1CSQ (bottom) as examples of good agreement with the LS model. Residues marked in white correspond to D_{KL} values greater than 0.02.

Table captions

Tab.1. List of analyzed proteins; particularly fast-folding proteins, cold shock proteins and prefoldins. PDB IDs (italicized) represent fast-folding proteins. The name of the protein is listed in second column. Protein lengths (expressed by the number of residues) are also given. The first number (where two numbers are listed) expresses the fragment (domain) taken for fast-folding process analysis, according to the reference given in the rightmost column. The second number indicates the full length of the protein. A simplified description of biological function is listed in the following column. The final column lists papers which describe the properties of the protein.

Tab. 2. List of non-redundant proteins taken for analysis.

Tab. 3. Sequence similarity and RMS-D for proteins with significant structural differences along with high sequence similarity.

Tab. 4. List of selected fast-folding proteins. $D_{average}$ describes the mean distance between the expected and measured value of the radius of curvature observed for a particular V-angle. $D_{average}$ is taken as the criterion of consistency with the ES model. PDB IDs of fast-folding proteins are italicized.

O/T and O/R express the distance entropy (Kullback-Leibler) used to compare the observed (O) hydrophobicity distribution with idealized values (T) and with random distribution (R). These parameters are then used as criteria of consistency in the LS model. O/T < O/R is considered to imply agreement between the idealized and observed hydrophobicity distributions. The rightmost column lists structural characteristics (H – helix, B – β-structure, R – random, D – dimer, L – ligand, DR – DNA/RNA) where the protein molecule is involved in interaction with other molecules.

The values for proteins recognized as accordant with assumed models are highlighted.

Tab. 5. Summary of the ES and LS accordance between the observed protein structure and the assumed model. Identifiers of fast-folding proteins are italicized.

Table caption – Supplementary materials

Tab.1S. Sequence similarity (expressed in % to show the broad spectrum of sequence differentiation in the proteins taken for analysis).

REFERENCES

Aboderin A. (1971) An empirical hydrophobicity scale for alpha-amino-acids and some of its applications. Int. J. Biochem. 2: 537-544.

Banach M, Stapor K, Roterman I. (2009) Chaperonin structure – the large multi-subunit protein complex. Int. J. Mol. Sci. 10: 844-861.

Banach M, Prymula K, Jurkowski W, Konieczny L, Roterman I. (2011) Fuzzy oil drop model to interpret the structure of antifreeze proteins and their mutants. J. Mol. Model. PMID: 21523554.

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, 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.

Bogatyreva NS, Osypov AA, Ivankov DN. (2009) Kinetic DB: a database of protein folding kinetics. Nucleic Acids Res. 37: D342–D346.

Brylinski M, Konieczny L, Roterman I. (2006) Hydrophobic collapse in (in silico) protein folding. Comput. Biol. Chem. 30(4):255-67.

Bryliński M, Kochańczyk M, Broniatowska E, Roterman I. (2007a) Localization of ligand binding site in proteins identified in silico. J. Mol. Model. 13: 655-675.

Brylinski M, Konieczny L, Roterman I. (2007b) Is the protein folding an aim-oriented process? Human haemoglobin as example. Int. J. Bioinform. Res. Appl. 3(2): 234-60.

Chen C, Chen L, Zou X, Cai P. (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein & Peptide Letters 16: 27-31.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acid Res. 31(13):3497-3500.

Cho JH, O'Connell N, Raleigh DP, Palmer AG. 3rd. (2010) Phi-value analysis for ultrafast folding proteins by NMR relaxation dispersion. J. Am. Chem. Soc. 132(2):450-451.

Chou KC. (1989) Graphic rules in steady and non-steady enzyme kinetics. J Biol. Chem. 264: 12074-12079.

Chou KC. (1990) Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys. Chem. 35: 1-24.

Chou KC. (1993) Graphic rule for non-steady-state enzyme kinetics and protein folding kinetics. J. Math. Chem. 12: 97-108.

Chou KC. (1995a) Does the folding type of a protein depend on its amino acid composition? FEBS Lett. 363: 127-131.

Chou KC. (1995b) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function & Genetics 21: 319-344.

Chou KC. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Structure, Function, and Genetics (Erratum: ibid., (2001) 44: 60) 43: 246-255.

Chou KC. (2010) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J Theor. Biol. 273(1): 236-47.

Chou KC, Carlacci L. (1991) Energetic approach to the folding of alpha/beta barrels. Proteins: Struct., Funct., Genet 9: 280-295.

Chou KC, Scheraga HA. (1982) Origin of the right-handed twist of beta-sheets of poly-L-valine chains. Proc. Nat. Acad. Sci. USA 79: 7047-7051.

Chou JJ, Zhang CT. (1993) A joint prediction of the folding types of 1490 human proteins from their genetic codons. J Theor. Biol. 161: 251-262.

Chou KC, Zhang CT. (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol. Chem. 269: 22014-22020.

Chou KC, Zhang CT. (1995) Review: Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30: 275-349.

Chou KC, Maggiora GM, Neméthy G, Scheraga HA. (1988) Energetics of the structure of the four-alpha-helix bundle in proteins. Proc. Nat. Acad. Sci. USA 85: 4295-4299.

Chou KC, Maggiora GM, Scheraga HA. (1992) Role of loop-helix interactions in stabilizing four-helix bundle proteins. Proc. Nat. Acad. Sci. USA 89: 7315-7319.

Chou KC, Neméthy G, Scheraga HA. (1983a). Role of interchain interactions in the stabilization of right-handed twist of ß-sheets. J Mol. Biol. 168: 389-407.

Chou KC, Neméthy G, Scheraga HA. (1983b) Effects of amino acid composition on the twist and the relative stability of parallel and antiparallel β-sheets. Biochemistry 22: 6213-6221.

Chou KC, Neméthy G, Scheraga HA. (1984) Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. Journal of American Chemical Society 106: 3161-3170.

Chou KC, Neméthy G, Scheraga HA. (1990) Review: Energetics of interactions of regular structural elements in proteins. Acc. Chem. Res. 23: 134-141.

Chou KC, Neméthy G, Rumsey S, Tuttle RW, Scheraga HA. (1986) Interactions between two beta-sheets: Energetics of beta/beta packing in proteins. J. Mol. Biol. 188: 641-649.

Chou KC, Shen HB. (2008) Cell-PLoc: A package of Web servers for predicting sub-cellular localization of proteins in various organisms. Nature Protocols 3: 153-162.

Chou KC, Shen HB. (2009a) Review: recent advances in developing web-servers for predicting protein attributes. Natural Science 2: 63-92 (freely accessible at http://www.scirp.org/journal/NS/).

Chou KC, Shen HB. (2009b) FoldRate: A web-server for predicting protein folding rates from primary sequence. The Open Bioinformatics Journal 3: 31-50 (freely accessible at http://www.bentham.org/open/tobioij/).

Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO. (1994) Structural studies of the engrailed homeodomain. Protein Sci. 3: 1779–1787.

Consortium U. (2009) The universal protein resource (UniProt) 2009. Nucleic Acids Res. 37: D169–D174.

Day R, Daggett V. (2003) All-atom simulations of protein folding and unfolding. Adv Protein Chem. 66: 373–403.

Dimitriadis G, Drysdale A, Myers JK, Arora P, Radford SE, Oas TG, Smith DA. (2004) Microsecond folding dynamics of the f13w g29a mutant of the B domain of staphylococcal protein A by laser-induced temperature jump. Proc. Natl. Acad. Sci. USA 101: 3809–3814.

Ding H, Luo L, and Lin H. (2009) Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein & Peptide Letters 16: 351-355.

Dyer RB. Ultrafast and downhill protein folding. (2007) Curr. Opin. Struct. Biol. 17: 38–47.

Englander SW. (2000) Protein folding intermediates and pathways studied by hydrogen exchange. Annu. Rev. Biophys. Biomol. Struct. 29: 213-238.

Falzone CJ, Mayer MR, Whiteman EL, Moore CD, Lecomte JT. (1996) Design challenges for hemoproteins: the solution structure of apocytochrome b5. Biochemistry 35: 6519–6526.

Fisher AC, DeLisa MP. (2008) Laboratory evolution of fast-folding green fluorescent protein using secretory pathway Quality control. PLoS ONE 3(6) e2351.

Ghosh K, Ozkan SB, Dill KA. (2007) The ultimate speed limit to protein folding is conformational searching. J. Am. Chem. Soc. 129: 11920–11927.

Guo HX, Rao NN, Liu GX, Li J, and Wang YH. (2010) Predicting protein folding rate from amino acid sequence. Progress in Biochemistry and Biophysics 37: 1331-1338

Hopfner KP, Kopetzki E, Kresse GB, Bode W, Huber R, Engh RA. (1998) New enzyme lineages by subdomain shuffling. Proc. Natl. Acad. Sci. USA 95: 9813-9818.

Jefferys BR, Kelley LA, Sternberg MJ. (2010) Protein folding requires crowd control in a simulated cell. J. Mol. Biol. 16: 1329-1338.

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

Kandaswamy KK, Chou KC, Martinetz T, Moller S, Suganthan PN, Sridharan S, and Pugalenthi G. (2011) AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270: 56-62.

Kauzmann W. (1959) Some factors in the interpretation of protein denaturation. Adv. Protein Chem. 14: 1–63.

Konieczny L, Brylinski M, Roterman I. (2006) Gauss-function-Based model of hydrophobicity density in proteins. In Silico Biol. 6: 15-22.

Kremer W, Schuler B, Harrieder S, Geyer M, Gronwald W, Welker C, Jaenicke R, Kalbitzer HR. (2001) Solution NMR structure of the cold-shock protein from the hyperthermophilic bacterium Thermotoga maritima. Eur. J. Biochem. 268: 2527-2539.

Kubelka J, Eaton WA, Hofrichter J. (2003) Experimental tests of villin subdomain folding simulations. J. Mol. Biol. 329: 625–630.

Kubelka J, Hofrichter J, Eaton WA. (2004) The protein folding 'speed limit'. Curr. Opin. Struct. Biol. 14: 76–88.

Kuhlman B, Luisi DL, Evans PA, Raleigh DP. (1998) Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the n-terminal domain of the protein l9. J. Mol. Biol. 284: 1661–1670.

Levitt M, (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 104 (1): 59–107.

Lin H, and Ding H. (2011) Predicting ion channels and their types by thedipeptide mode of pseudo amino acid composition. J. Theor. Biol. 269: 64-69.

Liu Y, Liu Z, Androphy E, Chen J, Baleja JD. (2004) Design and characterization of helical peptides that inhibit the E6 protein of papillomavirus. Biochemistry 43: 7421-7431.

Liu T, and Jia C. (2010) A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J. Theor. Biol. 267: 272-275.

Luisi DL, Kuhlman B, Sideras K, Evans PA, Raleigh DP. (1999) Effects of varying the local propensity to form secondary structure on the stability and folding kinetics of a rapid folding mixed alpha/beta protein: characterization of a truncation mutant of the N-terminal domain of the ribosomal protein L9. J. Mol. Biol. 289: 167-174.

Macias MJ, Gervais V, Civera C, Oschkinat H. (2000) Structural analysis of WW domains and design of a ww prototype. Nat. Struct. Biol. 7: 375–379.

Manyusa S, Whitford D. (1999) Defining folding and unfolding reactions of apocytochrome b5 using equilibrium and kinetic fluorescence measurements. Biochemistry 38: 9533–9540.

Mao B, Chou KC, and Zhang CT. (1994) Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins. Protein Eng. 7:319-330.

Masso M, and Vaisman II. (2010) Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. J. Theor. Biol. 266:560-568.

Mayor U, Johnson CM, Daggett V, Fersht AR. (2000) Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. Proc. Natl. Acad. Sci. USA 97: 13518–13522.

Max KE, Zeeb M, Bienert R, Balbach J, Heinemann U. (2007) Common mode of DNA binding to cold shock domains. Crystal structure of hexathymidine bound to the domain-swapped form of a major cold shock protein from Bacillus caldolyticus. Febs J. 274: 1265-1279.

Mohabatkar H. (2010) Prediction of Cyclin Proteins Using Chou's Pseudo Amino Acid Composition. Protein & Peptide Letters 17:1207-1214.

Nalewajski RF. (2006) Information theory of molecular systems. Amsterdam [etc.]: Elsevier, ISBN 978-0-444-51966-5.

Nguyen H, Jager M, Moretto A, Gruebele M, Kelly JW. (2003) Tuning the free-energy landscape of a ww domain by temperature, mutation, and truncation. Proc. Natl. Acad. Sci. USA 100: 3948–3953.

Ohtaki A, Kida H, Miyata Y, Ide N, Yonezawa A, Arakawa T, Iizuka R, Noguchi K, Kita A, Odaka M, Miki K, Yohda M. (2008) Structure and molecular dynamics simulation of archaeal prefoldin: the molecular mechanism for binding and recognition of nonnative substrate proteins. J. Mol. Biol. 376: 1130-1141.

Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. (1999) Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins. Suppl. 3:149-70.

Ozkan SB, Dill K, Bahar I. (2002) Fast-folding protein kinetics, hidden intermediates and the sequential stabilization model Protein Science 11: 1958-1970.

Pande VS, Grosberg AY, Tanaka T, Rokhsar DS. (1998) Pathways for protein folding: Is the new view needed? Curr. Opin. Struct. Biol. 8(1): 68-79.

Perl D, Walker C, Schindler T, Schröder K, Marahiel MA, Jaenicke R, Schmid FX. (1998) Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. Nat. Struct. Biol. 5: 229–235.

Prymula K, Sałapa K, Roterman I. (2010) "Fuzzy oil drop" model applied to individual small proteins built of 70 amino acids. J. Mol. Model. 16: 1269-1282.

Prymula K, Piwowar M, Kochanczyk M, Flis L, Malawski M, Szepieniec T, Evangelista G, Minervini G, Polticelli F, Wiśniowski Z, Sałapa K, Matczyńska E, Roterman I. (2009) In silico structural study of random amino acid sequence proteins not present in Nature Chemistry and Biodiversity 6: 2311-2336.

Roterman I. (1995) Modelling the optimal simulation path in the peptide chain folding--studies based on geometry of alanine heptapeptide. J. Theor. Biol. 177: 283-288.

Roterman I, Bryliński M, Konieczny L, Jurkowski W. (2007) Early-stage protein folding – In silico model. In: Structural Bioinformatics. Ed. Alexandre de Brevern, Research Signpost, Kerala India.

Schindelin H, Marahiel MA,Heinemann U.(1993) Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. Nature 364: 164-168.

Schindelin H, Jiang W, Inouye M, Heinemann U. (1994) Crystal structure of CspA, the major cold shock protein of Escherichia coli. Proc. Natl. Acad. Sci. USA 91: 5119-5123.

Shen HB, Song JN, and Chou KC. (2009) Prediction of protein folding rates from primary sequence by fusing multiple sequential features. Journal of Biomedical Science and Engineering 2:136-143 (freely accessible at http://www.scirp.org/journal/jbise)

Spector S, Raleigh DP. (1999) Submillisecond folding of the peripheral subunit-binding domain. J. Mol. Biol. 293: 763–768.

Stayrook SE, Jaru-Ampornpan P, Ni J, Hochschild A, Lewis M. (2008) Crystal structure of the lambda repressor and a model for pairwise cooperative operator binding. Nature 452: 1022-1025.

Wallace R. (2010) Protein folding disorders: toward a basic biological paradigm. J. Theor. Biol. 267:582-594.

Wang T, Zhu Y, Gai F. (2004) Folding of a three-helix bundle at the folding speed limit. The Journal of Physical Chemistry B 108: 3694–3697.

Wang JF, Chou KC. (2009) Insight into the molecular switch mechanism of human Rab5a from molecular dynamics simulations. Biochem. Biophys. Res. Commun. 390:608-612.

Zakeri P, Moshiri B, and Sadeghi M. (2011) Prediction of protein submitochondria locations based on data fusion of various features of sequences. J. Theor. Biol. 269:208-216.

Zeeb M, Max KE, Weininger U, Low C, Sticht H, Balbach J. (2006) Recognition of T-rich single-stranded DNA by the cold shock protein Bs-CspB in solution. Nucleic Acids Res. 34: 4561-4571.

Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, Li ML. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259:366-372.

Zhang CT, and Chou KC. (1992) Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. Biophys. J. 63:1523-1529.

Zhu Y, Alonso DO, Maki K, Huang CY, Lahr SJ, Daggett V, Roder H, DeGrado WF, Gai F. (2003) Ultrafast folding of alpha3d: a de novo designed three-helix bundle protein. Proc. Natl. Acad. Sci. USA 100: 15486–15491.

Zobnina V, Roterman I. (2009) Application of the fuzzy-oil-drop model to membrane protein simulation Proteins Structure, Function, Bioinformatics 77: 378-394.

PDB ID	Protein/domain name	Length	Biological function	Ref.
1VII_A	Villin subdomain	826/36	protein binding	Kubelka et al 2003 Kubelka et al 2004
1E0L_A	WW domain FBP 28	1100/37	cs-trans isomerase	Nguyen et al 2003 Macias et al 2000
1IET_A	Cytochrome b5	134/94	heme binding	Spector and Raleigh 1999 Falzone et al 1996 Manyusa and Whitford 1999
2PDD_A	Oxidoreductase	428/43	protein binding	Spector and Raleigh 1999
1PRB_A	Albumin binding domain	387/53	protein binding	Wang et al 2004
1CQU_A	50S ribosomal protein L9	149/56	rRNA binding	Kuhlman et al 1998 Luisi et al 1999
1BDD_A	Protein A, B domain	508/60	IgG binding	Dimitriadis et al 2004
2A3D_A	A3D	73	de novo designed	Zhu et al 2003
1CSQ	Cold Shock protein	67	nucleic acid binding	Schindelin et al. 1993
1YPA_I	Subtilisin inhibitor 2A	84/64	inhibitor	Day and Daggett 2003
1FXY_1	Coagulation factor	228/107	coagulation factor	Hopfner et al 1998
1G6P	Cold shock protein	66	nucleic acid binding	Kremer et al 2001

1MJC	Cold Shock protein	69	nucleic acid binding	Schindelin et al 1994
1RIJ	E6	23	binding seven-residue leucine-containing motif.	Liu et al 2004
2HAX	Cold Shock protein	66	nucleic acid binding	Max et al. 2007
2ZDI	Prefoldin	101	chaperone	Ohtaki et al 2008
3BDN	Lambda repressor	236	DNA binding	Stayrook et al 2008
1CQU	Ribosomal L9 protein	56	ribosomal prot.	Luisi et al. 1999

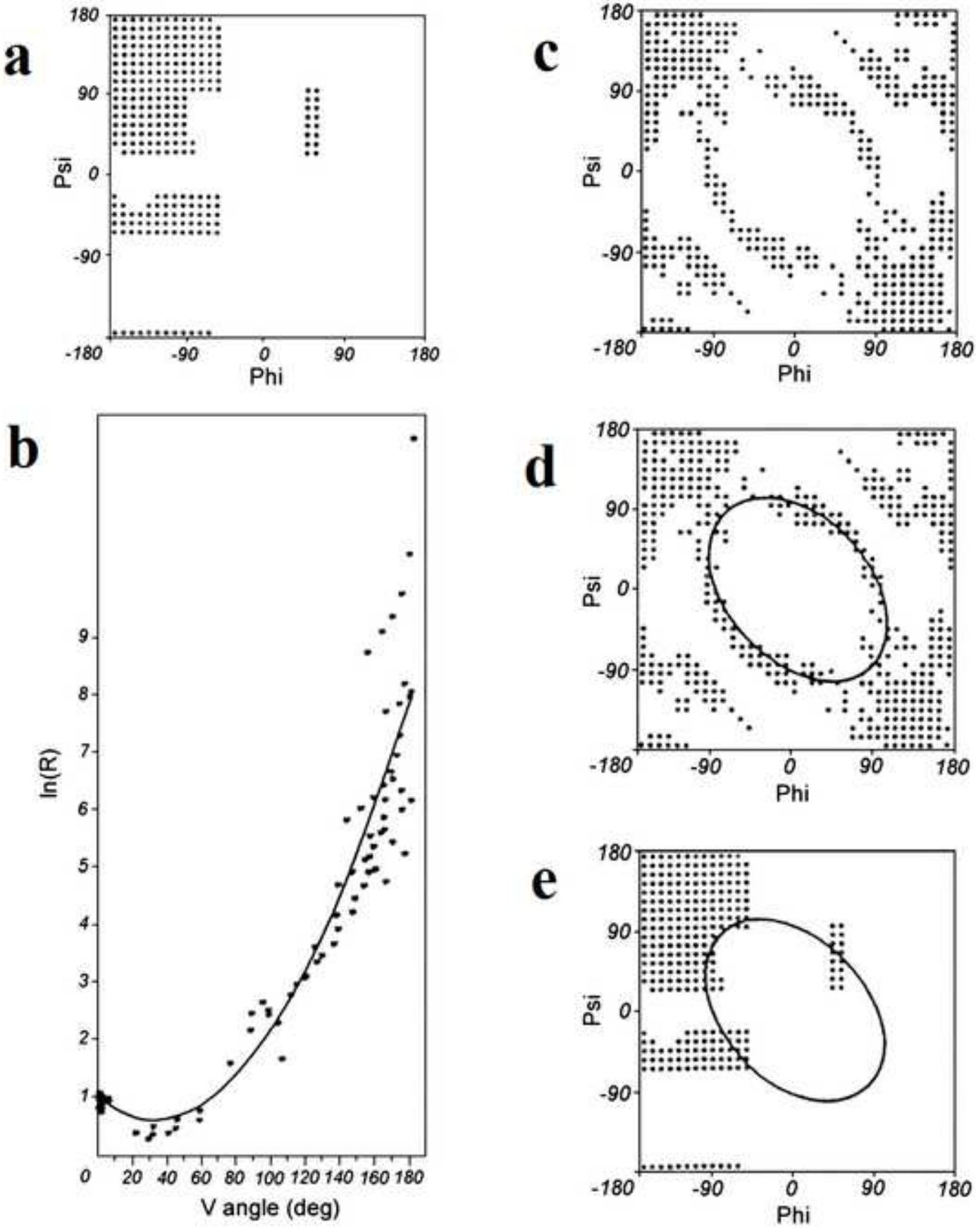
NON-REDUNTANT PROTEINS
1BDD_A
1CQU_A
1E0L_A
1FXY_A
1G6P_A
1IET_A
1MJC_A
1PRB_A
1RIJ_A
1VII_A
1YPA_I
2A3D_A
2HAX_A
2PDD_A
2ZDI_A
3BDN_A

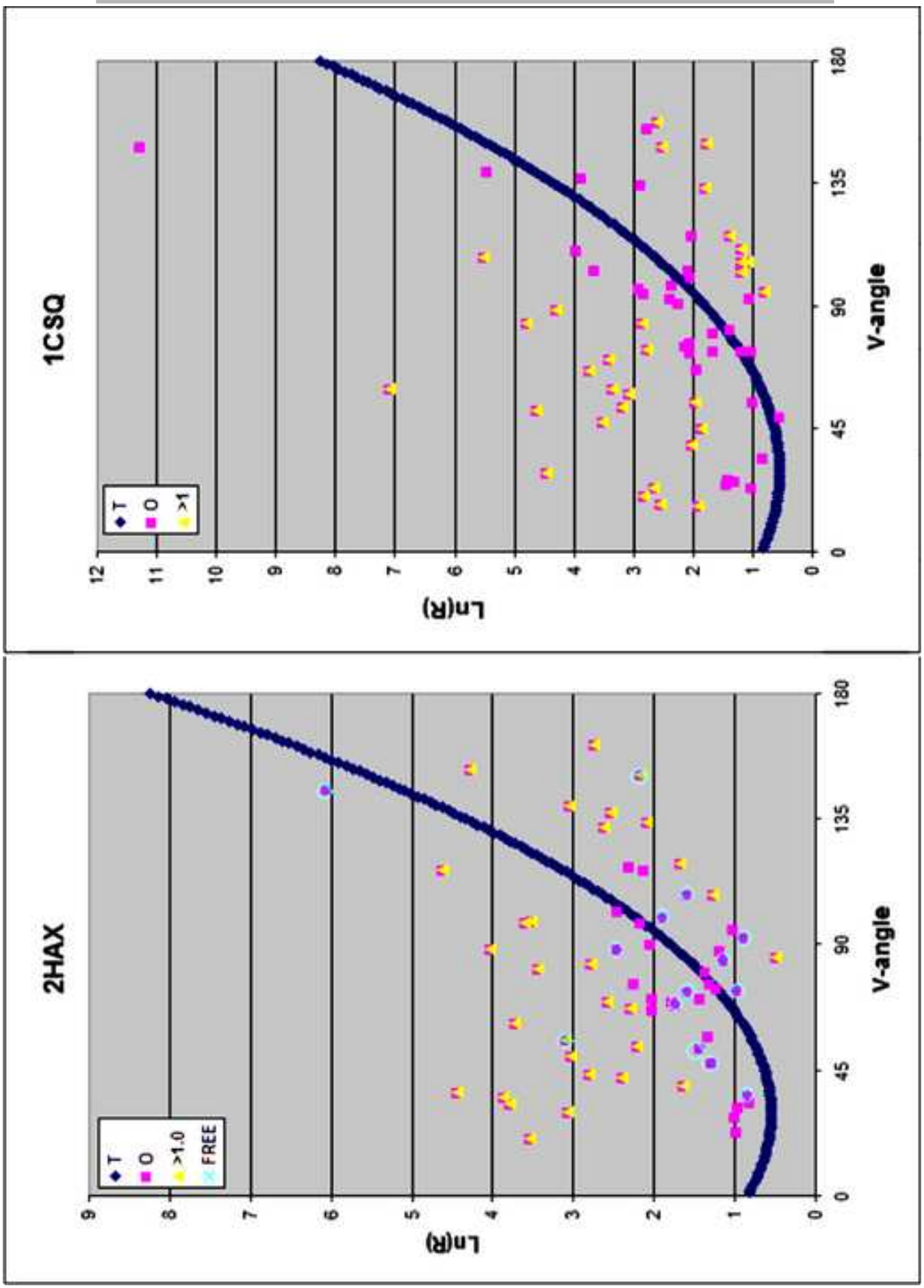
PROTEIN	#AA	PROTEIN	#AA	SEQ %	RMS-D
1G6P_A	66	2HAX_A	66	62	18.36
1RIJ_A	23	3BDN_A	236	26	4.85
1PRB_A	53	2ZDI_A	117	22	17.11
1VII_A	36	2ZDI_A	117	22	10.05
1YPA_I	64	2ZDI_A	151	20	7.13

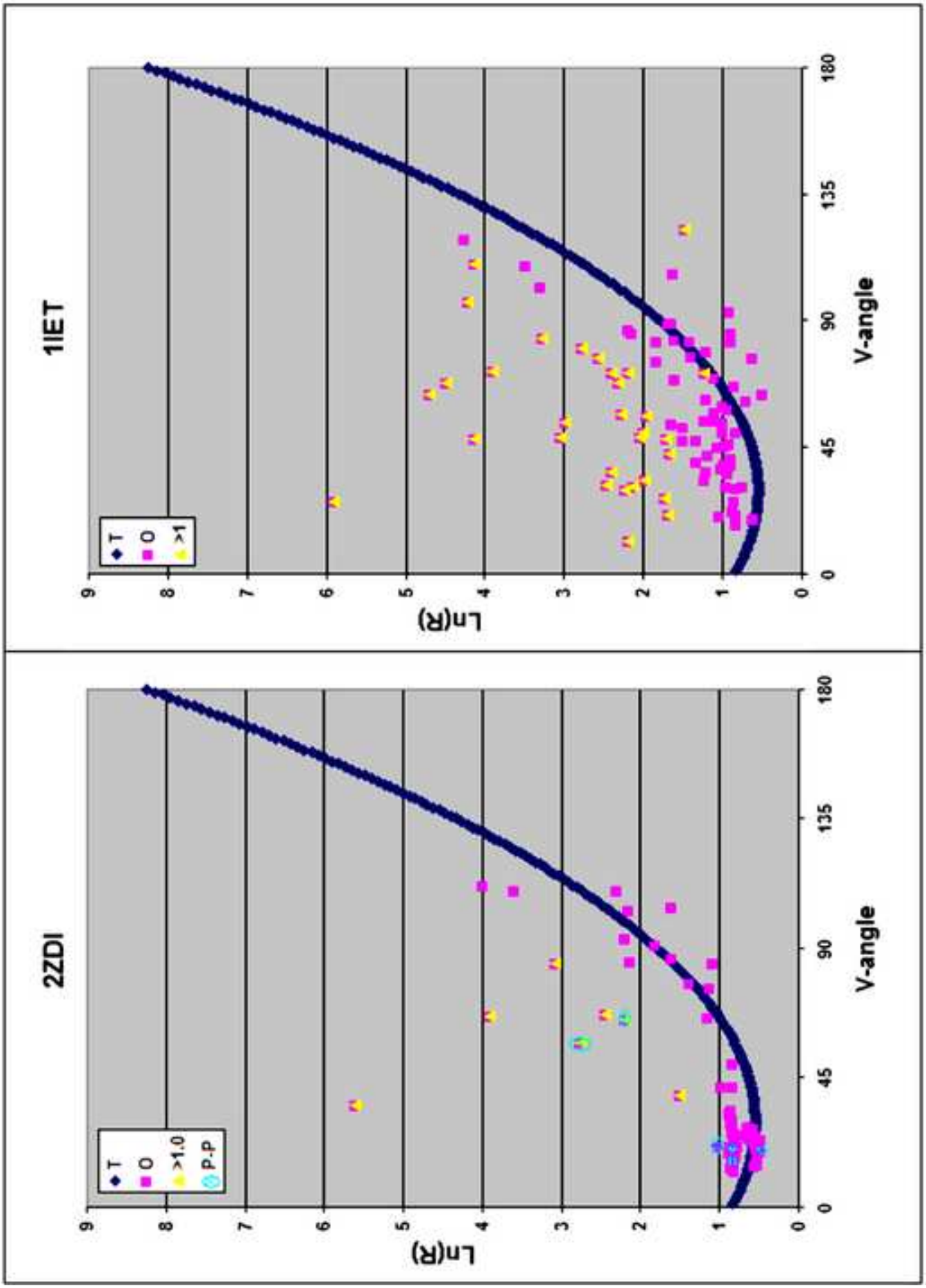
PDB ID	N	$D_{average}$	O/T	O/R	Structure characteristics
1IET_A	94	**0.901**	**0.2706**	**0.5855**	H + B + R
2PDD_A	43	**0.521**	**0.1694**	**0.4609**	H + R
1YPA_I	83/64	**0.812**	**0.1088**	**0.3286**	H + B + R
1BDD_A	60	**0.371**	**0.1860**	**0.5855**	H + R
1PRB_A	55	**0.306**	**0.2462**	**0.5798**	H
1VII_A	75/36	**0.275**	**0.2232**	**0.5677**	H + R
1E0L_A	37	1.010	**0.1419**	**0.2976**	B + R
2A3D_A	73	**0.323**	**0.3719**	**0.5153**	H + R
1CQU_A	57	**0.792**	**0.3327**	**0.6856**	H + B + R
1FXY	107	**0.351**	**0.1355**	**0.2243**	β –Barrel
1G6P	66	1.348	**0.0923**	**0.2949**	β –Barrel
1MJC	69	1.254	**0.0909**	**0.3915**	β -barrel
1RIJ	23	**0.581**	**0.1715**	**0.5832**	H + R
1CSQ	67	1.600	**0.1696**	**0.2036**	β -barrel
2HAX_A	66	1.283	0.6290	0.4034	B+D+L+DR
2ZDI_A	101	**0.366**	0.5922	0.4549	H + D
3BDN	236	**0.984**	0.5976	0.3607	H + R

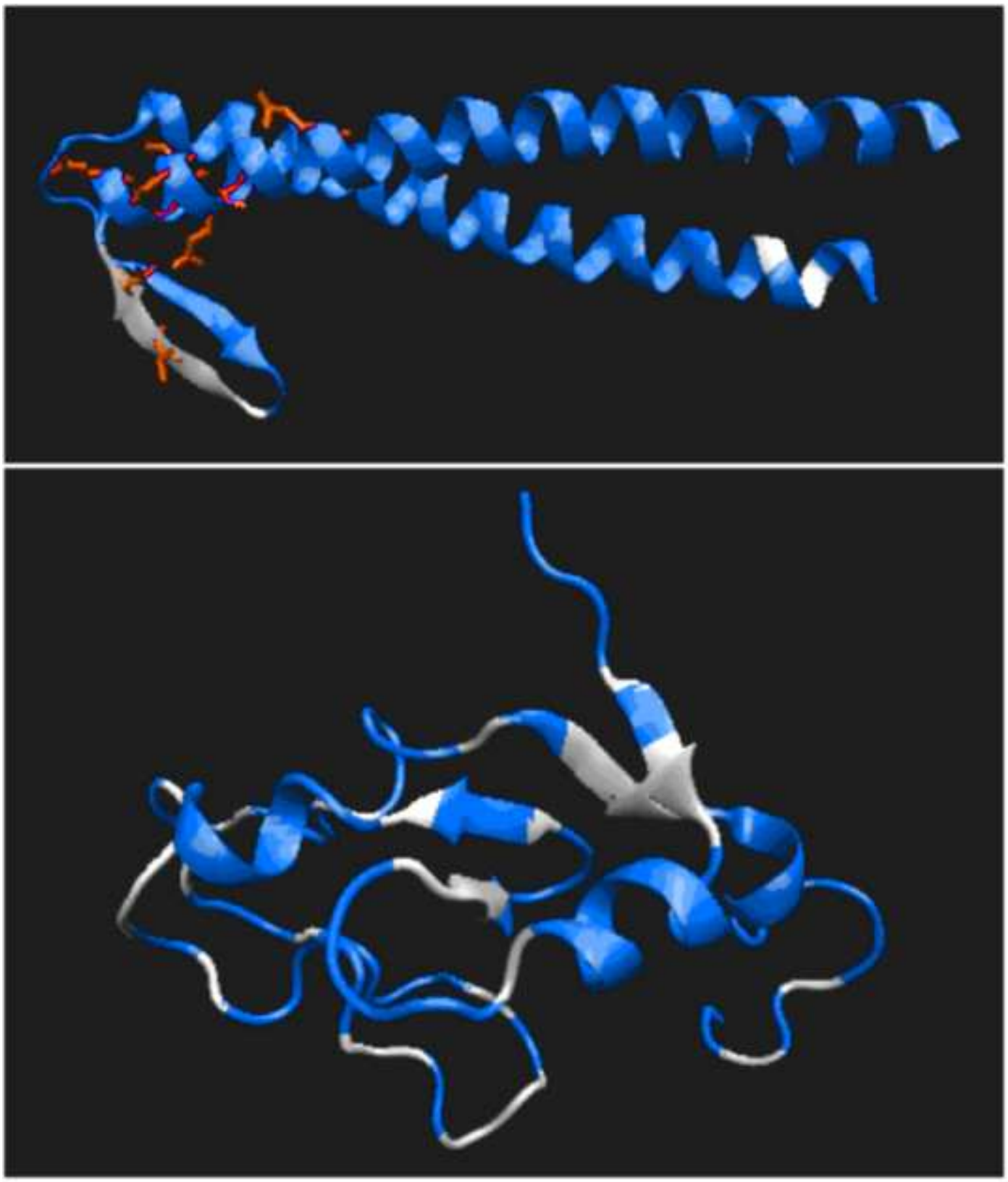
MODEL APPLIED		ES MODEL	
		ACCORDANT	NOT ACCORDANT
LS MODEL	ACCORDANT	*1IET, 2PDD, 1YPA, 1BDD, 1PRB, 1VII, 2A3D, 1CQU,* 1FXY, 1RIJ	*1E0L,* 1MJC, 1G6P, 1CSQ
	NOT ACCORDANT	2ZDI, 3BDN	2HAX

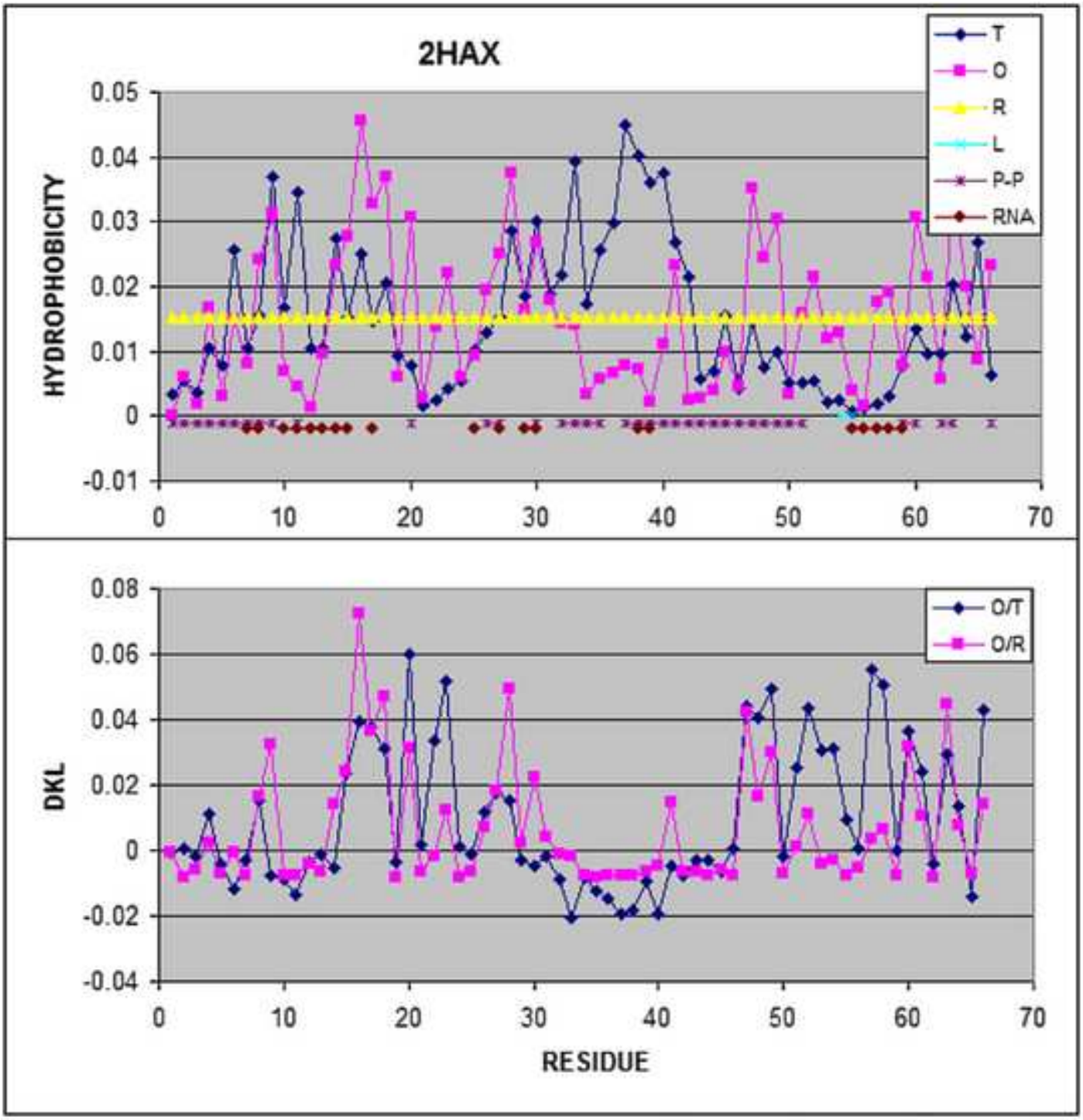
Two-step process for protein folding process is presented using down-hill proteins. > The accordance with assumed early and late stage intermediate is presented. > The early-stage intermediate is analyzed using backbone conformation. > The late-stage intermediate is characterised by hydrophobic core structure. > The accordance is measured using elements of information theory.

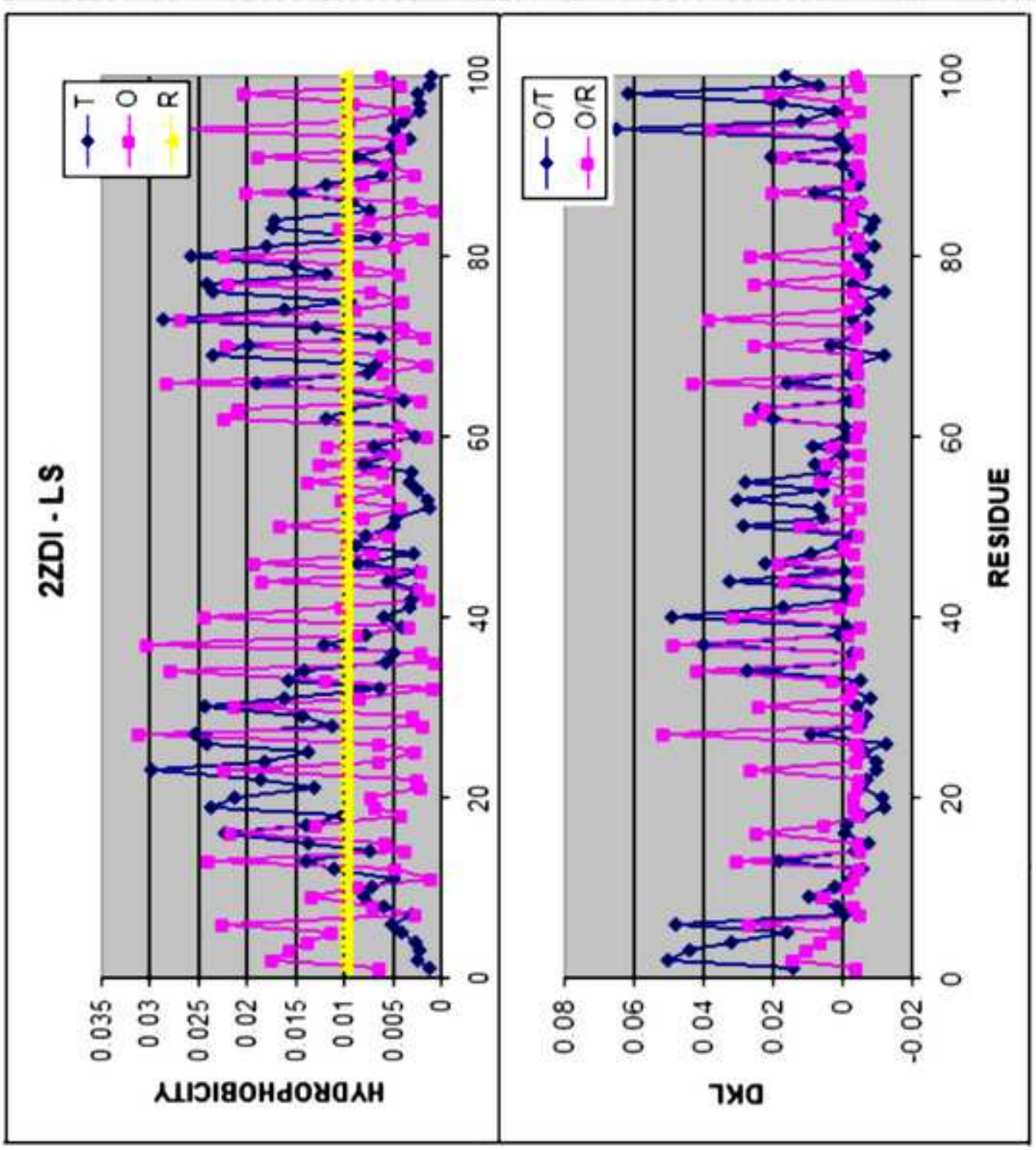


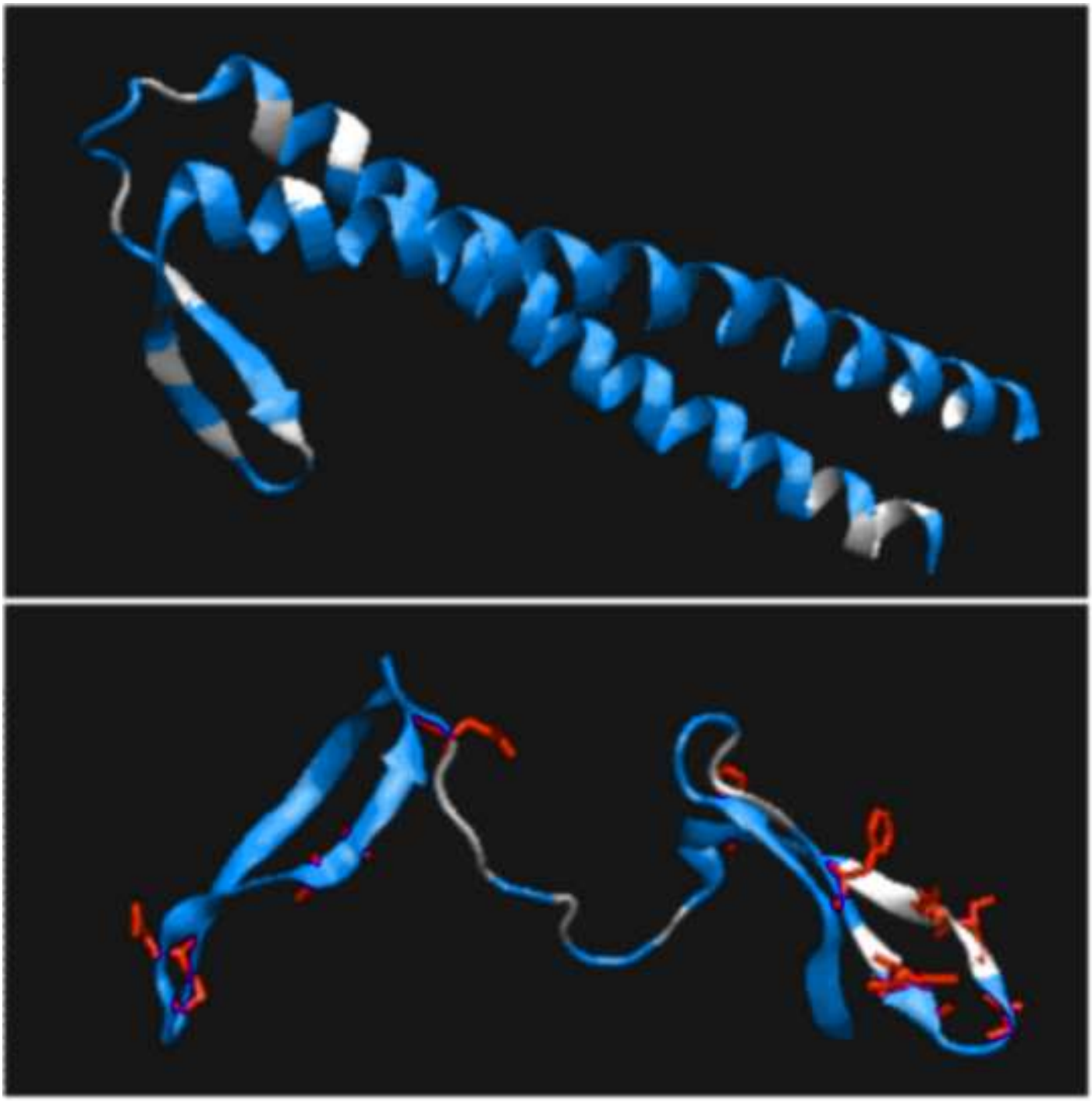

1CSQ
Ln(R)
V-angle
T
O
>1
2HAX
Ln(R)
V-angle
T
O
>1.0
FREE

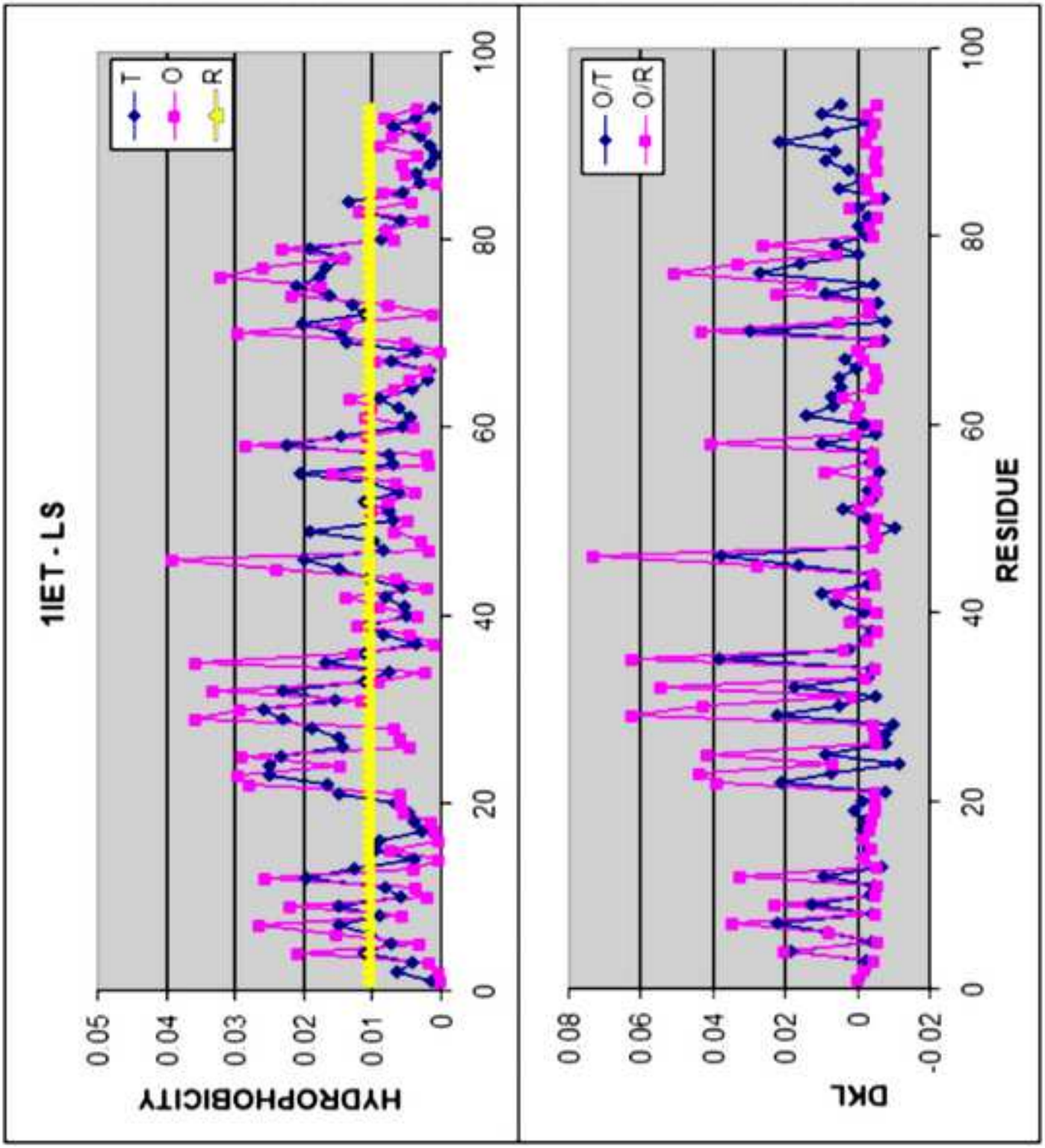


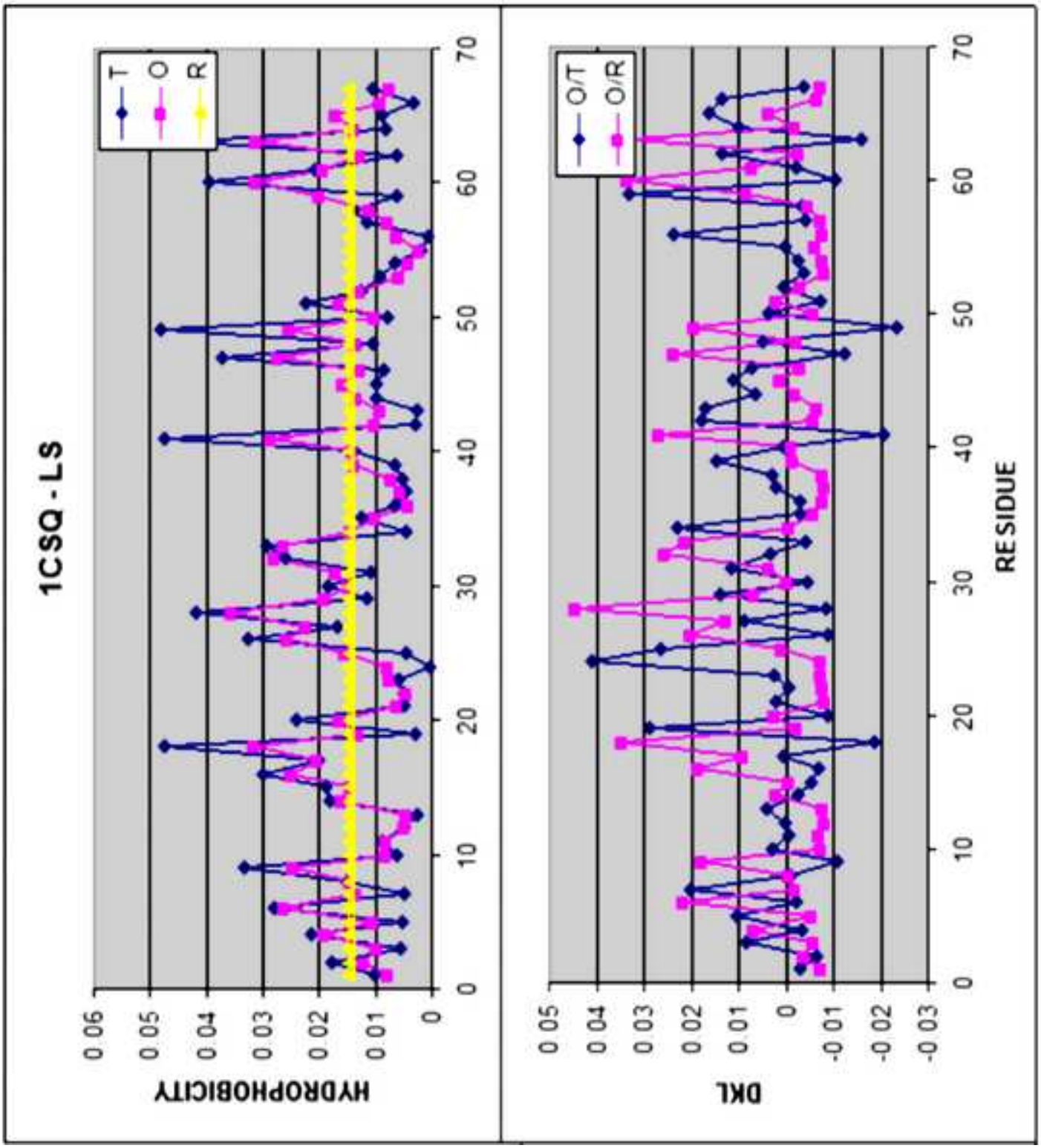





2ZDI - LS
T
O
R
HYDROPHOBICITY
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
O/T
O/R
DKL
0.08
0.06
0.04
0.02
0
-0.02
RESIDUE




1IET - LS
HYDROPHOBICITY
T
O
R
0.05
0.04
0.03
0.02
0.01
0
0
20
40
60
80
100
DKL
O/T
O/R
0.08
0.06
0.04
0.02
0
-0.02
RESIDUE


1CSQ - LS
HYDROPHOBICITY
T
O
R
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50
60
70
DKL
O/T
O/R
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
RESIDUE