Molecular BioSystems

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

Cite this: *Mol. BioSyst.*, 2013, **9**, 1774

Network properties of decoys and CASP predicted models: a comparison with native protein structures†

S. Chatterjee,‡[a] S. Ghosh‡[ab] and S. Vishveshwara*[a]

Received 18th April 2013,
Accepted 13th May 2013

DOI: 10.1039/c3mb70157c

www.rsc.org/molecularbiosystems

Protein structure space is believed to consist of a finite set of discrete folds, unlike the protein sequence space which is astronomically large, indicating that proteins from the available sequence space are likely to adopt one of the many folds already observed. In spite of extensive sequence–structure correlation data, protein structure prediction still remains an open question with researchers having tried different approaches (experimental as well as computational). One of the challenges of protein structure prediction is to identify the native protein structures from a milieu of decoys/models. In this work, a rigorous investigation of Protein Structure Networks (PSNs) has been performed to detect native structures from decoys/models. Ninety four parameters obtained from network studies have been optimally combined with Support Vector Machines (SVM) to derive a general metric to distinguish decoys/models from the native protein structures with an accuracy of 94.11%. Recently, for the first time in the literature we had shown that PSN has the capability to distinguish native proteins from decoys. A major difference between the present work and the previous study is to explore the transition profiles at different strengths of non-covalent interactions and SVM has indeed identified this as an important parameter. Additionally, the SVM trained algorithm is also applied to the recent CASP10 predicted models. The novelty of the network approach is that it is based on general network properties of native protein structures and that a given model can be assessed independent of any reference structure. Thus, the approach presented in this paper can be valuable in validating the predicted structures. A web-server has been developed for this purpose and is freely available at http://vishgraph.mbu.iisc.ernet.in/GraProStr/PSN-QA.html.

Introduction

Proteins are known to adopt unique well-defined three-dimensional structures to carry out their functions efficiently.[1] The number of available protein sequences far exceeds that of the solved structures, emphasizing the need to model the structures and thereby understand the sequence–structure–function relationships in proteins. Predicting the three-dimensional structure of a native protein, given its amino acid sequence, has been a challenge for many decades. Towards this goal, many large scale statistical studies on experimentally determined protein structures have been performed, each describing various aspects that are crucial to reconstruct a well-folded protein structure. Such studies have highlighted the roles of specific amino acids as helix breakers or initiators.[2] Further, the role of hydrophobic residues to form the core of a protein structure and the various pair-wise interactions that are crucial to impart stability to secondary and super-secondary structures have also been investigated.[3–8] Based on these large scale statistical studies, gross understandings of the rules that govern protein folding have been accomplished and are widely discussed in the literature.[9–12]

For a number of years now, protein structures have been treated as networks that help in obtaining a mathematical abstraction of the global topological features.[13,14] Based on this approach investigation such as identification of groups of interacting residues important for folding/function, residues important for protein stability and fluctuations[15,16] have been performed. Graph theoretical concepts have been used to address the problem of protein structure selection,[17] and further to discriminate native proteins from their decoy sets.[18,19] In the majority of these studies, the protein structure networks are generated at the C_α/backbone level, which although useful, may not be sufficient to capture the chemical and geometrical properties that arise due to side chain interactions of a residue. Studies on protein structure networks

[a] *Molecular Biophysics Unit, Indian Institute of Science, Bangalore – 560012, India. E-mail: sv@mbu.iisc.ernet.in*

[b] *I.I.Sc. Mathematics Initiative, Indian Institute of Science, Bangalore – 560012, India*

† Electronic supplementary information (ESI) available: See DOI: 10.1039/c3mb70157c

‡ Both authors contributed equally to the work.

This journal is © The Royal Society of Chemistry 2013

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

[PSNs] that are generated at the level of side chain atoms have been extensively carried out in our laboratory and it is observed that adding side chain details makes the network much more realistic, allowing the network to capture minute details of the protein topology.[20–22] Sidechain based PSN studies of native proteins show patterns that are missing in random networks, highlighting that paradigms do exist in nature that govern protein folding and also underlines the capability of PSNs to capture the same.[23]

One of the approaches to understand protein folding is to identify parameters for the comparison of native and decoy/model structures. Decoys are protein structures with minor conformational deviations from their native fold and are generated by various computational methods, such as molecular dynamics simulations[24–26] and discrete state models.[27] It is believed that the root mean square deviation (RMSD), hydrogen bond patterns, accessible surface area, interaction energy of amino acid pairs[28–30] and the position of conserved residues[31] in decoy/model sets differ from the native structures, therefore forming an excellent source to study and identify unique properties of a natively folded protein structure. Another critical aspect of modelling protein structure is to assess the quality of structures generated. This involves developing effective methods and designing scoring functions to rank the structures based on their quality. Studies have been performed to develop these scoring functions based on accessible surface area and amino acid neighbourhood considerations.[32] Many knowledge based potentials[33,34] have also been developed based on large scale studies to identify interactions that are preferred in a natively folded structure. Characterising native structures using the network properties of a protein structure are also available in the literature.[17–19] A large scale community-wide prediction of protein structures and evaluation of the predicted models have been facilitated by the Critical Assessment of Structure Prediction [CASP] group[35] and forms an excellent resource to obtain non-native like structures.

As mentioned above, our laboratory has focused on PSNs arising due to the interaction of side-chains. Recently, we explored the potential of PSNs to distinguish decoys/models from the native protein structures[36] for the first time in the area of protein structure prediction. The network properties calculated for PSNs constructed at lower interaction strength [(I_{min}) = 1%, described in the method section] indicated that such a formalism can be used to distinguish the native structures from decoys. On the other hand, the transition profile of network parameters as a function of interaction strength has been shown to be unique for the native protein structures[37] and an exploration of network features at all levels of interaction strengths would increase this distinguishing capability. In the present study these features are incorporated in a rigorous manner. Towards this goal, the PSNs are constructed and the packing of residues at different I_{min}s are explored to study the changes in network parameters, as it transitions from a dense network (lower I_{min}) to a sparse network (higher I_{min}). While, the PSNs capture network properties at the side chain level and are much more detailed, main chain hydrogen bonds [MHB] have also been shown to be crucial for packing of the polypeptide chain[8] and have been included in this study. These features that characterises native proteins are passed to Support Vector Machines (SVM)[38] to generate a classifier. SVMs have been widely used for predicting protein secondary structures,[39–42] protein SCOP classes[43] and protein binding sites.[44] Machine learning algorithms have also been used for predicting the quality of protein structures[45,46] using features based on secondary structures, solvent accessibility and atom–atom contacts. In the present study, the network properties of the PSNs are used to build a SVM classifier.

In the present study, a total of 94 features that best describes the various aspects of a PSN are integrated to build an SVM model, which shows an overall accuracy of 94.11%. The features are also ranked based on how well they distinguish the decoy/model structures from the natives. Interestingly, the transition behaviour of the protein as it progresses from lower I_{min} to higher I_{min} has been shown to contribute maximally to distinguish the decoy from the native structures, highlighting the uniqueness of transition profiles of network parameters in the native structures. The model is further validated using near-native decoy structures from the Rosetta dataset,[47,48] quality assessed structures of CASP9 [http://predictioncenter.org/casp9] and native structures [www.pdb.org]. A complete network analysis of the recently released CASP10 models has also been performed and compared with CASP10 assessments. Furthermore, a free web-server (http://vishgraph.mbu.iisc.ernet.in/GraProStr/PSN-QA.html) has also been developed to check the quality of a model. Details of the methodology are provided in the next section, followed by results and discussion.

Methods

Dataset selection

Decoy/model datasets are downloaded from the following sources and structures with residues lower than hundred are filtered out to ensure a proper hydrophobic core:

(a) University of Libre de Bruxelles (http://babylone.ulb.ac.be/decoys)[29]

(b) Baker laboratory (http://depts.washington.edu/bakerpg/decoys/)[47,48]

(c) Decoys 'R' Us (http://dd.compbio.washington.edu/download.shtml)[49]

(d) CASP3/7/8/9/10 (http://predictioncenter.org/casp3/ http://predictioncenter.org/casp7/ http://predictioncenter.org/casp8 http://predictioncenter.org/casp9 http://predictioncenter.org/casp10)

A total of 308 native structures and their corresponding decoys/models (29 543) are obtained from the above sites. Additionally, PDB structures are also downloaded from RCSB[50] (www.pdb.org) to increase the representation of native structures in the datasets. The monomeric protein crystal structures with resolution less than 3 Å, *R*-factor less than 0.25 and residue number greater than 100 are selected. Finally, the native dataset consists of 5422 proteins and the decoy/model set consists of 29 543 structures. Individual details for each dataset are provided in Table 1 and details of the individual proteins in the dataset are provided in Table S1 (ESI†).

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

Table 1 Total number of native and decoy/model structures from each dataset used in the present study. The last column lists the website from which the datasets are downloaded

Dataset	Number of native proteins	Number of decoy proteins	Website
CASP3	1	971	http://predictioncenter.org/download_area/CASP3/
CASP7	6	10	http://predictioncenter.org/download_area/CASP7/
CASP8	95	10 299	http://predictioncenter.org/download_area/CASP8/
CASP9	53	7711	http://predictioncenter.org/download_area/CASP9/
CASP10[b]	29	1428	http://predictioncenter.org/download_area/CASP10/
Rosetta protein decoy set	19	2660	http://depts.washington.edu/bakerpg/decoys/
Standard and complete collection of decoy set	5	1799	http://babylone.ulb.ac.be/decoys
Single decoy set	17	17	http://dd.compbio.washington.edu/download.shtml
Haemoglobin structural set	21	609	http://dd.compbio.washington.edu/download.shtml
Immunoglobulin structural set	61	3659	http://dd.compbio.washington.edu/download.shtml
Immunoglobulin structural hire set[a]	20	380	http://dd.compbio.washington.edu/download.shtml

[a] The native proteins are a subset of the immunoglobulin structural set; however the number of decoys and the method of generation are different. [b] As of 11th September 2012.

Construction of protein structure network [PSN]

A Protein Structure Network [PSN] gives an insight into the topological features of a protein structure based on the non-covalent side-chain interactions. Details to build a protein structure network are provided in previous studies[20,37] and a brief description is provided here. Amino acids in the protein structures are considered as nodes and edges are made between residues with non-covalent interactions and quantified by the parameter $[I_{ij}]$:

$$I_{ij} = \frac{n_{ij} \times 100}{\sqrt{N_i \times N_j}} \quad (1)$$

where, I_{ij} = strength of interaction between residues i and j, where $|i - j| \geq 2$; n_{ij} = number of distinct interacting atom pairs between i and j within a distance cut-off of 4.5 Å (excluding the backbone atoms); N_i and N_j are the normalization values for residues i and j obtained from a statistically significant dataset of proteins, as defined earlier.[20] PSNs are constructed at different interaction strengths as defined by I_{min} and the values ranging from 0% to 7% have been used in this study. The network parameters which capture the global topological features such as, simple pair-wise interactions [NCov], size of the largest cluster [SLClu] (largest cluster is calculated using the Depth First Search algorithm[51]), size of the largest k-2 community [ComSk2],[52] cumulative size of the top3 k-1 community [Top3-ComSk1],[53] clustering coefficients of the PSN [CCoe], sub-network formed by the largest cluster [CCoe-LClu] and largest k-2 community [CCoe-Lcomm] have been calculated and analysed for the different categories. Differences in the values of the aforementioned global parameters at successive I_{min}s are calculated for a discrete representation of the slope of the transition profiles. Table 2 provides the network parameters used in the present study with a brief description of each parameter. Fig. 1 shows the representation of k-1 and k-2 community on an example protein structure (1ERV).

Generation of random network models

The behaviour of the network properties is also compared with random networks. Two extreme cases of random networks,[23] RM1 and RM3, are generated. The RM1 models are based on the Erdos–Renyi random networks and have no similarity to protein structure topology. The size of the network is fixed to 100 and networks are generated at different probability values that correspond to the connections obtained at different I_{min} values. The RM3 models are specific cases of random models where the position of each node/residue and the number of interactions between the residues are kept the same as in native PSN at a given I_{min} and the interactions [edges] are then randomly generated throughout the PSNs. Ten random networks at different I_{min} for RM1 and RM3 are generated.

Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm, used mainly for dataset classification. This method uses a training set, which consists of instance-label pairs (attributes and class label for a set of data points) to obtain a hyperplane on the nth dimension that best separates the data. The rules learnt from the training set are used to build a classifier which is further validated using test cases.[38] LIBSVM[54] is a library of SVM that is freely available (http://www.csie.ntu.edu.tw/~cjlin/libsvm) and is used in the present study.

The network parameters provided in Table 2 are calculated at different I_{min} [from 0% to 7%] for the native and the decoy/model datasets. Along with these, MHB for each native and decoy/model structure is calculated using HBPlus,[55] with the default settings. Finally, a total of 94 features are obtained to build the SVM classifier. While most of the parameters are derived from the network properties of the PSNs at different I_{min} and describe the features at the side chain level one of the parameters, MHB, describes the structure at the back bone level.

Since the number of decoy/model structures far exceeds the number of native structures, the dataset is distributed randomly 10 times, such that in each random subset, the training set consists of an equal number [3000] of native proteins and decoy/model structures, while the remaining are considered as a test set. Data pre-processing is performed in two steps. Firstly, the data is scaled to range from −10 to 10. Since, the features provided in SVM are independent of each other and the scale of

 This journal is © The Royal Society of Chemistry 2013

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

Table 2 List of the various network properties used in this study, accompanied by a one line description for each parameter

Parameter	Description
NCov	Number of non-covalent interactions, defined by the number of edges in a PSN
SLClu	Set of connected nodes with maximum number of residues (evaluated using DFS algorithm)[51]
Top1-ComSk1[a]	A clique is a subset of nodes in the network, such that all nodes are connected to all other nodes. Union of k-cliques such that k-1 nodes are shared between the cliques are termed as k-1-community.[53] This parameter represents the size of the largest k-1-community
Top2-ComSk1[a]	Cumulative size of the top2 largest k-1-community
Top3-ComSk1[a]	Cumulative size of the top3 largest k-1-community
ComSk2[a]	Union of k-cliques such that k-2 nodes are termed as k-2-community.[52] Represents the size of the largest k-2-community. Fig. 1 represents the k-2 community as well as a k-1 community in a protein structure
CCoe[b]	Avg. clustering coefficient of the network, based on the algorithm given in ref. 63
CCoe-LClu	Avg. clustering coefficient of the largest cluster. This was calculated by extracting the subnetwork that forms the largest cluster
CCoe-Lcomm	Avg. clustering coefficient of the largest k-2 community
d(NCov)#	Represents the transition profile of non-covalent interaction as a function of I_{min}
d(SLClu)#	Represents the transition of the size of the largest cluster as a function of I_{min}
d(ComSk2)#	Represents the transition of the size of the largest k-2 community as a function of I_{min}

[a] k-1 community and input for calculating k-2 community is obtained using CFinder.[53] [b] All the parameters except those marked with '#' are normalised by the size of the protein.

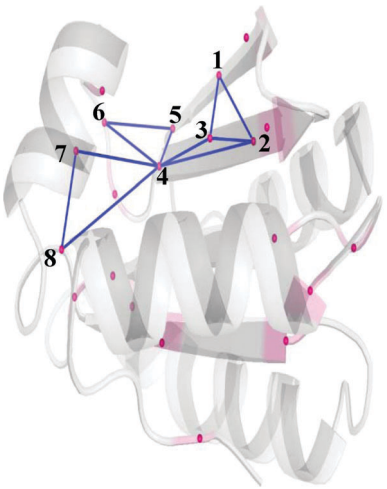

Fig. 1 Community is a union of cliques formed by sharing a minimum of (k-2) nodes represented on the protein [1ERV at I_{min} = 1%]. Participating nodes in the community are shown as pink spheres and the part of the connections having four (k = 3) cliques are shown as blue lines. The sub-community [k-1 community] is formed by the nodes (1, 2, 3, 4) by sharing the edge (2, 3). The cliques (2, 3, 4) (4, 5, 6) (4, 7, 8) share a common node 4 forming a k-2 community.

the value for different features may vary, the data is scaled such that in all the features the lowest value = −10 and the highest value = 10. On doing this, any bias that arises due to the different value ranges of the different features are removed. The radial basis kernel function (RBF) has been used, since the number of features is less than the number of data points and so a non-linear kernel function is required to map the data on a higher dimensional space. Using this, iterative cross-validation accuracy is performed to obtain the optimal values of c and g that give the best accuracy for the given training set. 'c' defines the strictness of the classification, with higher 'c' values implying increased strictness. The parameter g (gamma), defines the smoothness of the boundary that separates the two classes, with a larger gamma value representing a smoother and regular boundary.

Table 3 Accuracy and AUC (area under curve) values for 10 random subsets, where the training set contains 3000 native structures and 3000 decoy/model structures

Random subset	Accuracy (%)	AUC
d1	93.68	0.9854
d2	94.11	0.9853
d3	93.52	0.9843
d4	93.73	0.9847
d5	93.51	0.9847
d6	93.64	0.9854
d7	93.83	0.9855
d8	93.68	0.9851
d9	93.95	0.9850
d10	93.55	0.9853

A good classifier would be one that has higher c and g values. However, a trade-off does exist as higher strictness can lead to over-classification of data making the classifier less robust and general. Once the dataset is optimised, classification accuracy is calculated based on the LIBSVM protocol. Table 3 gives the details of accuracy obtained for the 10 random datasets. The dataset [d2] with the best accuracy is considered for further analysis.

Feature selection

Feature selection is a method commonly applied in SVM to identify features that best classify the data. Identifying important features not only helps in understanding the basis of classification but also reduces run time by removing the unimportant features from the SVM classification model. Feature selection is performed based on two methods, each of which is described below.

(a) Fselect:[56] calculates the F-score and measures the discrimination capability of a feature. The following equation is used to calculate the F-score

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (2)$$

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

where, $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ = average of the i^{th} feature in all, positive and negative datasets; $x_{k,i}^{(+)}$, $x_{k,i}^{(-)}$ = i^{th} feature of the k^{th} data point in positive and negative datasets respectively. A higher *F*-score implies the greater discrimination capability of the i^{th} feature. A limitation of this method is that it does not provide mutual information among any two features. Despite this, the method is simple and quite effective. Python implementation of Fselect is freely available at the LIBSVM site and is used to calculate scores for each feature.

(b) ReliefF:[57] this is another method for estimating attribute importance. Unlike Fselect, this method does not assume that the features are independent of each other and therefore are capable of identifying best features even if the features are dependent on each other. It uses regression analysis to calculate the importance of each feature. Matlab implementation of ReliefF is used for the present study.

Validation of the model

The classifier obtained is further validated using the near native structures of Rosetta, quality assessed structures of CASP9, and all 5422 native proteins. The near native decoy structures of Rosetta serve as the negative control, while all native proteins form the positive control. The CASP9 quality assessed structures are the best modelled structures and consist of a mixture of native like and near native structures.

CASP10 predictions

While the manuscript was under preparation, native structures for the CASP10 targets were not completely released. Native structures for the CASP10 targets released on or before 11th September 2012 were included in the training set and test set to build and test the classifier. However, the complete list of predicted models and the corresponding native structures (barring a few) were released later and have been analysed using the SVM classifier and critically compared with the CASP10 quality assessment.

Results

The network parameters described in Table 2 are evaluated for each of the 5422 natives and 29 543 decoy/model structures at different I_{min}. Additionally, the back-bone packing feature is captured by calculating the MHB for all the structures. Behaviour of the network features of all the decoys/models and natives, results pertaining to the classifier, validation and prediction are presented in the following sections.

Network features of the native and decoy/modelled structures

As explained in the method section, PSNs are generated for each of the protein native and decoy/model structures and the network parameters evaluated. The list of parameters is provided in Table 2 and a brief description is provided here. NCov represents the number of non-covalent interactions in a protein structure and is a good measure of the integrity of the protein structure. Since side chain atoms are the major contributors of non-covalent interactions, this parameter also accounts for the chemistry of the amino acid in a structural context. Size of the largest cluster [SLClu] represents the global connectivity of the protein structure. A cluster is defined as the set of connected nodes and the cluster that is made of the maximum number of nodes is termed as the largest cluster. This property represents bond percolation in a PSN. Other higher order connectivities such as cliques and communities are also calculated using CFinder.[53] A k-clique is defined as a set of *k* nodes, in which all nodes are connected to each other. A union of k-cliques formed by sharing k-1 or k-2 are called a k-1 or k-2 community respectively. A detailed representation of communities is shown in Fig. 1. Large communities represent the percolation of highly connected units. Both k-1 and k-2 communities are evaluated for the present study. The clustering coefficient [CCoe] is a measure of the inter-connectivity of nodes and defines the cliquishness of a PSN. Network behaviour is investigated for all the native and decoy/model sets through these parameters, by plotting them as a function of I_{min}. Below we discuss the nature of these parameters by presenting the results for CASP3, CASP7 and all the data put together.

Fig. 2 describes the transition profile of the four basic network properties [NCov, SLClu, ComSk2, CCoe] as a function of I_{min} for two example datasets, CASP3 [C3], CASP7 [C7] and the average property over all datasets [AD]. Fig. 2a shows the transition profile of NCov in a native protein and their decoys/models for all the three datasets. C3 consists of 971 modelled structures corresponding to a single native structure; while C7 consists of 6 native structures and AD consists of 5422 native structures. A closer look at Fig. 2a for C3 clearly suggests that the native profile [shown in blue] shows higher values at lower I_{min} and drastically falls down as it approaches higher I_{min}, when compared to its decoy structures. A similar pattern is observed for all the other network properties for C3 (SLClu, ComSk2, CCoe) and clearly indicates patterns specific to the natives and the decoys/models. If we observe the profiles for CASP9 and CASP10 datasets (Fig. S1, ESI†), a reverse trend is observed for NCov and SLClu at lower I_{min}, although the values are lower than that of the decoy structures at higher I_{min}. [It should be noted that plots for a single protein and its decoy can provide better insight as in the case of C3 while some information is lost when averaged over a large number of data points]. Nevertheless, the striking feature exhibited consistently by the network parameters is the transition of the values from I_{min} = 1% to I_{min} = 4%. This is highly evident from the middle panel of Fig. 2. This panel shows the transition profiles for C7 native structures but fails to show any significant transition for the decoy/model sets. This behaviour holds true for all the network properties and is also observed in the combined profile comparison for all the native proteins and decoys/models in the dataset (Fig. 2 right panel). This is further validated by comparing the profiles of native, decoy/model and random structures as shown in Fig. S2 (ESI†). Random structures have been earlier explored in our laboratory[23] in comparison to native structures to understand the percolation behaviour of a natively folded structure. Two extreme cases of random structures, RM1 and RM3, are considered for this analysis.

 This journal is © The Royal Society of Chemistry 2013

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

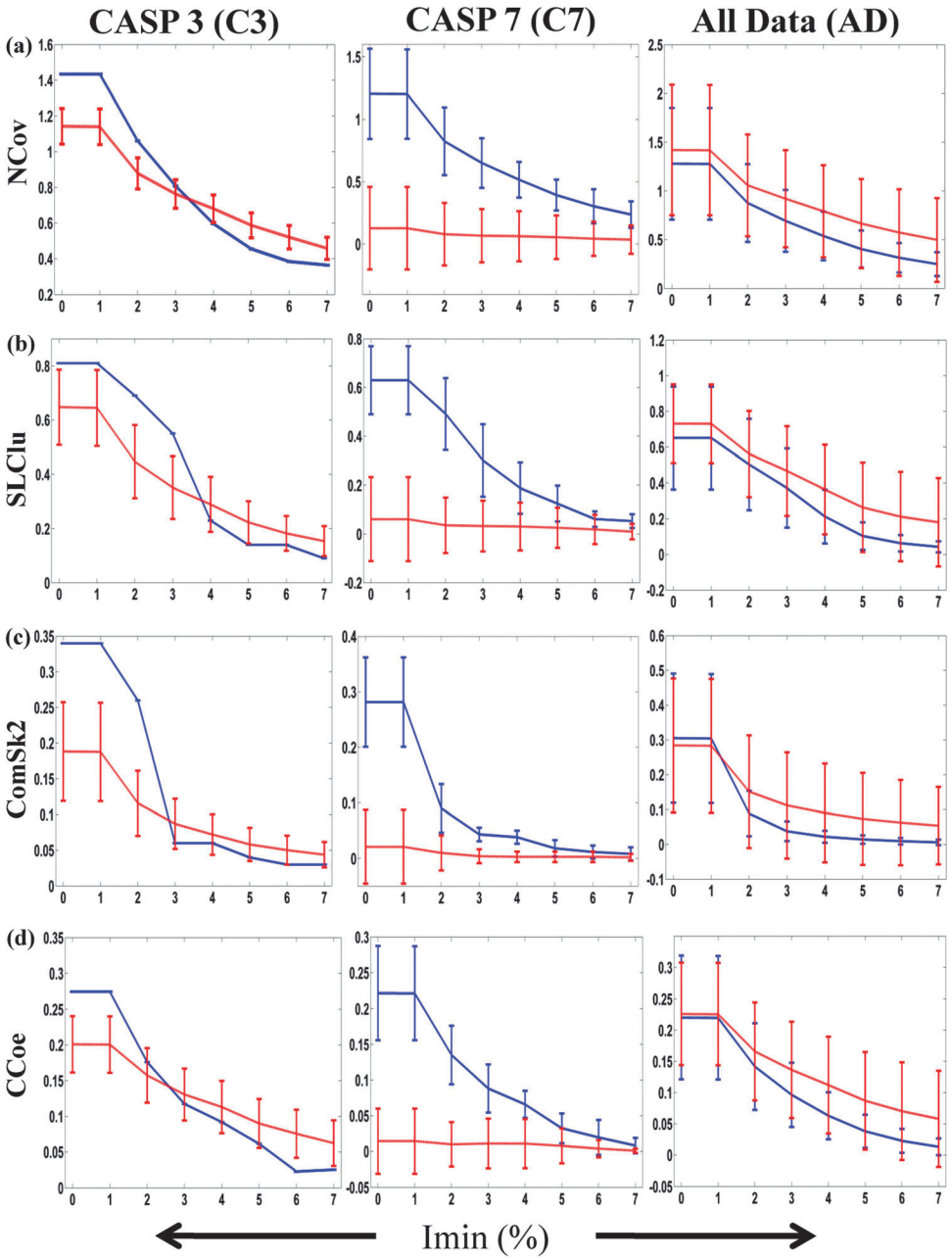

Fig. 2 Profiles of the four normalised network parameters, (a) NCov, (b) SLClu, (c) ComSk2 and (d) CCoe are shown for the three example datasets, (1) CASP3, (2) CASP7, (3) average over all datasets. Native structures are shown as blue lines while decoy/models are shown in red. The *X*-axis represents I_{min} from 0% to 7% and the *Y*-axis corresponds to the average value of the network property along with standard deviations.

For this study, the RM3 model was based on the CASP3 native structure (1BL0). As expected a distinct transition is completely missing from the profiles of RM1 structures (black lines) while the RM3 structures (green colour) show transition profiles nearer to the decoy/model structures than the native structures, emphasising the subtle but unique feature of the native structures.

The above results have shown that the interplay between the three network features (henceforth mentioned as characteristic features) (1) higher value at lower I_{min}, (2) lower value at higher I_{min} and finally (3) the transition from I_{min} = 1%–4% is able to capture the differences in the decoy/model and the native structures in a clear manner. As mentioned in the method section, the differences in the values of the parameters [such as d(NCov), d(SLClu) and d(ComSk2)] for the consecutive I_{min}s are calculated to represent this transitional behaviour. And these parameters are considered for building the SVM model along with other network parameters. Since, all the features represent the protein structure at the detailed side-chain level; other features that capture the backbone packing of the protein structure were also explored. Importance of the back-bone packing of protein structures has been discussed before.[8] We analysed the importance of MHB as a feature for SVM. MHB is

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

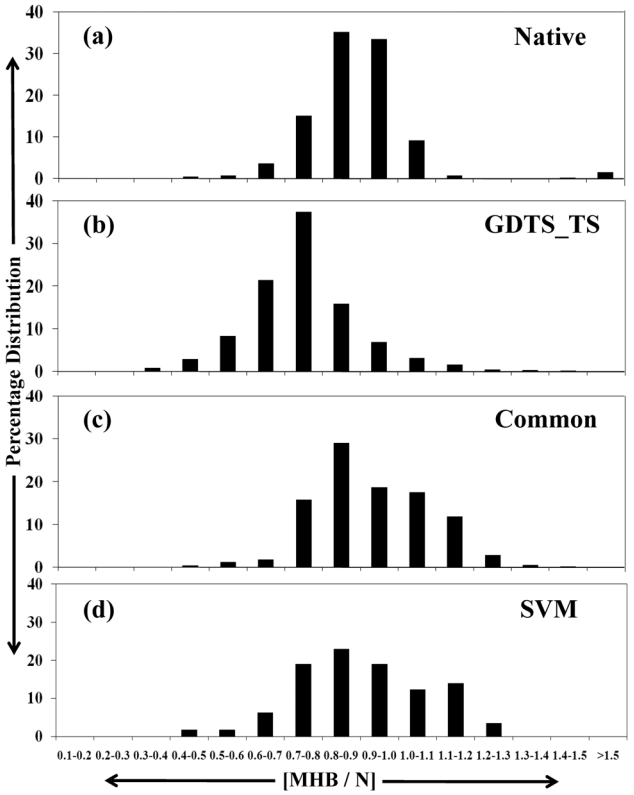

Fig. 3 Distribution plot of the main-chain hydrogen bond (normalised by the size of the protein) for (a) all the 5422 native proteins, (b) CASP10 models selected by only GDT-TS scores, (c) CASP10 models selected by GDT-TS and SVM and (d) CASP10 models selected by only SVM. The normalised value of MHB is shown along the *X*-axis, and the percentage of native protein is shown on the *Y*-axis.

calculated for all the 5422 native proteins using HBPlus[55] and a frequency plot generated. As can be observed from Fig. 3(a), the majority of the proteins showed an MHB [normalised by the size of the protein] value in the range of 0.6–1.1 (60% to 110% of the protein size). Therefore, the main-chain hydrogen bond is also included as a feature in SVM.

Support Vector Machine (SVM) classification

Here we have employed the techniques of SVM to understand the importance of topological parameters in determining the uniqueness of protein structures as described in the method section. Specifically, we are interested in distinguishing the native structures from the structures generated as decoys and identifying native and non-native structures from predicted models. The details of model building and the results pertaining to feature selection are described below.

The network properties described in the above section are integrated into a mathematical framework to build a model that best classifies the decoy/model structures from the native proteins, using SVM. 93 network features ranging from I_{min} = 0%–7% and MHB are integrated into the model. The data of 5422 native and 29 543 decoy/model structures are used as inputs to obtain ten random subsets. They are generated by randomly selecting 3000 native proteins and 3000 decoy/model structures as the training set and the remaining are used as test sets. The data is pre-processed before performing SVM as explained before. Table 3 lists the 10 subsets and the corresponding accuracy and AUC (area under curve) for each subset. Subset d2 gives the highest accuracy with an AUC = 0.9853 (Fig. 4a) and is considered for further analysis. It should however be noted that the other subsets have also yielded good accuracy.

Feature selection is used to identify features that can best differentiate the native structures from the decoy/modelled structures. This analysis helps in short-listing a set of features from the pool of available ones that truly captures the uniqueness of a protein structure. Two methods have been used for this analysis as described in the methods section. As can be seen from Fig. 4b, both the methods predict similar results, with MHB being the top scoring feature. Features that capture the transition profile of the protein structure as a function of I_{min}, such as d($NCov_{(2,3)}$), d($ComSk2_{(2,3)}$), d($SLClu_{(1,2)}$) are also predicted as top scorers along with SLClu, ComSk2, top1-ComSk1 and CCoe at different I_{min} values. Network properties such as NCov show similar differentiating capacity at all I_{min}s, while for parameters such as the SLClu, ComSk2, and CCoe, the differentiating capacity decreases at higher I_{min} values. This trend is reversed for the k-1 communities that can best capture the differences between the native structures and the decoy/model structures at higher I_{min} values. Table S2 (ESI†) gives a list of all the features used in the method and the corresponding *F*-scores and ReliefF scores. The results obtained by feature selection methods correspond very well with the patterns observed in Fig. 2 and again emphasises the interplay between the three characteristic features important for determining the quality of a native protein structure.

Model validation

During the process of model building the model is trained and tested using the respective datasets. Apart from these validations, some more rigorous tests are performed for specific cases. For this, we selected three datasets: (1) Rosetta near native structures, (2) CASP9 quality assessed structures and (3) all native structures.

Rosetta near native structures are known to be closer to the native structures and therefore differentiating these from the native structures is a challenging task. A dataset of Rosetta native structures (19) and their corresponding near native structures (380) are concatenated to form the input file. This forms a negative control and we expect the classifier to predict all the near native structures as decoy/model structures while all the native proteins as native. A striking accuracy of 99.74% is obtained. All the 380 near-native structures are classified as decoys, while 18 out of 19 structures are predicted as natives with the structure (4UBP), predicted as decoy. Fig. 5(a) plots the network parameters (SLClu, ComSk2 and CCoe) for native structures predicted as natives (blue), decoy structures predicted as decoys (green) and native predicted as decoy (red). If we look closely, 4UBP (red line) fails to show a significant

 This journal is © The Royal Society of Chemistry 2013

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

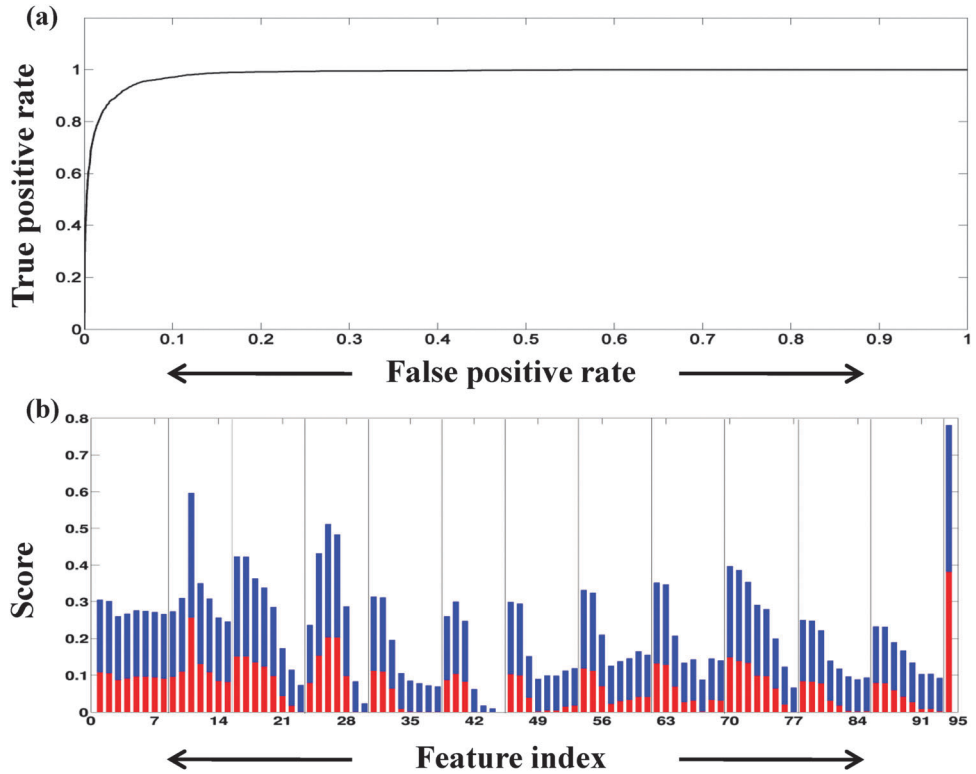

Fig. 4 SVM plots; (a) AUC plot based on different network features and main-chain hydrogen bond for the subset d2 exhibiting the highest accuracy of 94.11% (described in the method section). (b) A plot of *F*-score (red bar) and ReliefF score (blue bar) for the 94 features (*X*-axis) included in SVM. The actual values are presented in ESI† Table S2 and normalised values are plotted on the *Y*-axis.

transition for SLClu and also for other parameters at higher I_{min}. Also the values of ComSk2 and CCoe for 4UBP reduce to zero after the transition phase. Thus, the classifier is able to best capture the transition profile of network parameters as a function of I_{min} and successfully classify the structures.

CASP9 predicted structures are tested for quality by independent assessors and the best quality structures are labelled as quality assessed (QA). A set of quality assessment techniques have been developed over the years for different methods of structure prediction such as template-based,[58] free modelling based predictions,[59] residue–residue contact prediction[60] and so on. Details about these methods and various statistics used to rank the predictions are provided in previous articles.[35,61]

The SVM classifier built in this study is validated using the above-mentioned CASP9 quality assessed structures. In this case 54 native structures from CASP9 are concatenated with 1266 quality assessed CASP9 predictions and further classified using the classifier. Out of 54 native structures, 35 are predicted as native (true-positive) while 16 are predicted as non-native (false-negative) by the classifier. For the 1266 good-quality prediction models, 87 are predicted as native (false-positive) while the rest are predicted as decoy/model, giving an overall accuracy of 91.9%. To get a clear idea about this classification, we divided the structures into four classes: (1) native structures predicted as natives (nn), (2) native structures predicted as decoys/models (nd), (3) decoy/model structures predicted as decoys (dd) and finally (4) decoy/model structures predicted as natives (dn). The network profiles for each of the above classes are plotted as a function of I_{min} and a significant interplay of the three characteristic features is observed. If we look at the SLClu profile in Fig. 5(b), nn (blue) shows a higher value at lower I_{min} regions along with a steeper transition profile around I_{min} = 1%–4%. In fact, as expected the transition profile is steepest for nn as compared to dd (green). For the nd class (red), a lower value is observed at lower I_{min}. For the dn class (black), a smooth transition from lower to higher I_{min} is observed, however the values at lower I_{min} are lower than the native. A definite distinction can also be seen by studying the profiles for ComSk2 and CCoe (Fig. 5b). The nn class shows higher values at lower I_{min} and a steeper slope, while dd shows the reverse trend. It is interesting to note that the quality assessed structures being classified as dn class by SVM, have compensatory MHB values, as given in Table 4. The table provides the percentage distribution of MHB for the four classes. It is shown in Fig. 3(a) that the MHB for native structures mostly lie in the region of 0.6–1.1 (60–110% of the protein size). As can be seen for the dn class, the majority of the structures have MHB values in the favourable region and quite interestingly, most of the members of the nd class have values in the unfavourable region, highlighting the existing synergy between the properties of the side chain and the backbone. It seems that this synergy has been well picked up by the SVM model as also highlighted by the feature selection results. Amongst the network parameters, ComSk2 and CCoe quite distinctively pick up the differences between the native proteins and the decoy/modelled structures.

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

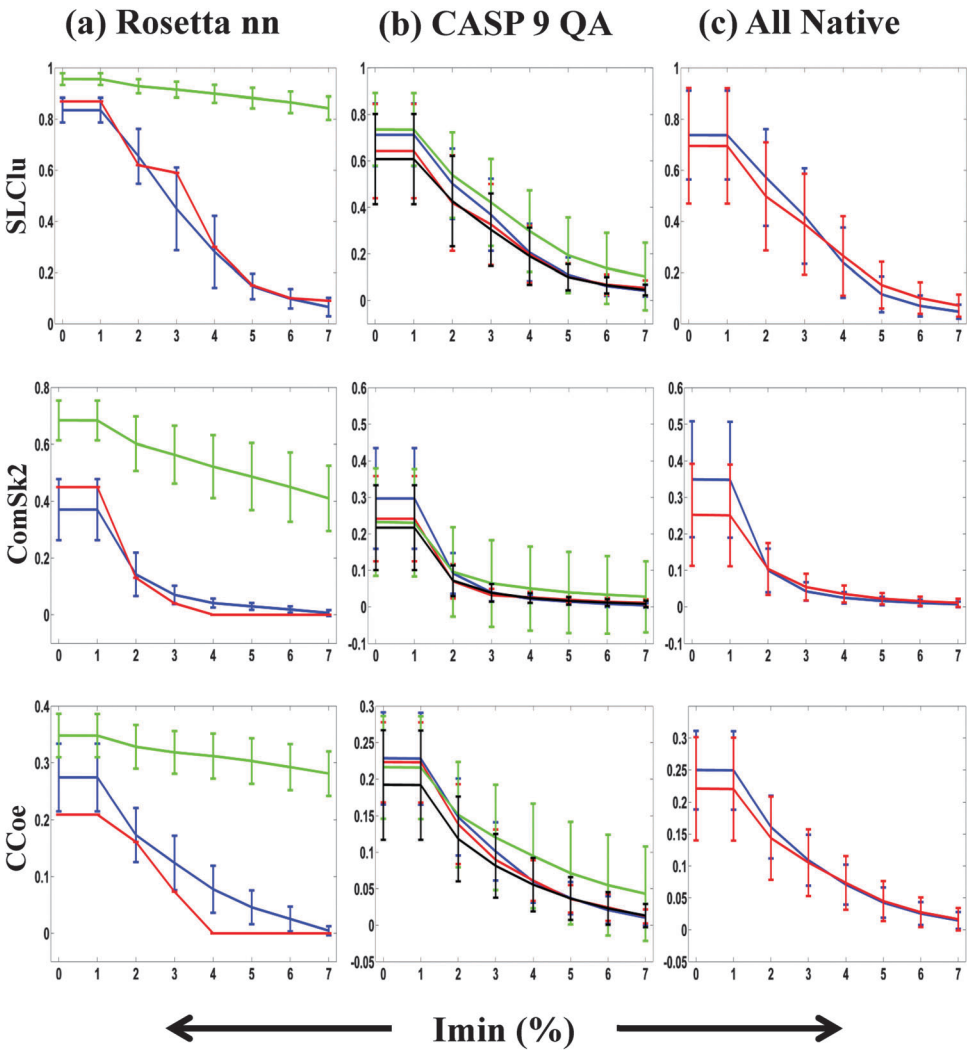

Fig. 5 SVM validation as shown by the profiles of the normalised network parameters (SLClu, ComSk2, and CCoe) as a function of I_{min} for (a) Rosetta-near native (19 native + 380 near-native), (b) CASP9 quality assessed structures (54 natives + 1266 models) and (c) 5422 native structures. Blue lines represent native structures predicted as native, green represent decoy/model structures predicted as decoys, red represent native structures predicted as decoys and black represent decoy/model structures predicted as natives as identified by SVM. I_{min} is plotted along the *X*-axis and the average values of the network properties along with standard deviations are shown on the *Y*-axis.

The third and final validation dataset is the 5422 native structures already downloaded from RCSB that served as a positive control. An overall accuracy of 95.11% was obtained with 5157 predicted as native and the rest (265) as decoy/model structures. Again the network properties of the true positives (nn) and false negatives (nd) are plotted as a function of I_{min} as shown in Fig. 5c. The difference in the network profiles for the two classes is clearly visible from the graph, with nn (blue) showing higher values at lower I_{min} and a steeper transition profile as compared to the other class nd (red). This behaviour is consistent in all the network properties, indicating the usefulness of this method.

CASP10 prediction

The SVM classifier is also used to assess the quality of the recently concluded CASP10 competition. A total of 31 869 predicted models corresponding to 83 targets are used [exclusive of NMR and cancelled targets], on which an RMSD filter of ≤ 5 Å is applied. This resulted in 8503 models of which 899 models are predicted as good quality from SVM. The CASP-10 site has provided GDT-TS[61,62] scores for all the models against which the results obtained by the SVM classifier are compared. The targets are divided as Template based modelling (TBM) or Free-modelling (FM) as given in CASP10. Out of the 83 targets, 54 are identified as TBM [8263 models] while 23 [240 models] are labelled as FM. In this study, statistics of CASP10 assessed scores and SVM selected models are shown for both TBM as well as FM targets; however, detailed analysis has been carried out only for TBM targets. Apart from the GDT-TS score, CASP10 predicted models are also assigned a side-chain specific, GDC score.[62] The SVM classifier is also compared against this score since both methods are based on side-chain atoms. A detailed discussion on the comparisons performed is provided below.

 This journal is © The Royal Society of Chemistry 2013

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

Table 4 Percentage MHB distribution of the four classes (nn, nd, dd, dn) identified by SVM for CASP9 quality assessed structures

	CASP9 quality assessed structure			
Range	nn	nd	dd	dn
0.1–0.2	0	0	0.424	0
0.2–0.3	0	0	1.613	0
0.3–0.4	2.632	25	2.716	0
0.4–0.5	23.68	62.5	7.555	0
0.5–0.6	7.895	12.5	13.67	4.598
0.6–0.7	2.632	0	24.02	9.195
0.7–0.8	0	0	25.72	16.09
0.8–0.9	7.895	0	16.04	36.78
0.9–1.0	5.263	0	6.197	13.79
1.0–1.1	7.895	0	1.104	12.64
1.1–1.2	0	0	0.509	4.598
1.2–1.3	5.263	0	0	1.149
1.3–1.4	0	0	0	1.149
1.4–1.5	2.632	0	0.424	0
>1.5	34.21	0	0	0

A broad distribution of the number of models that are selected by the different methods is presented in Fig. 6(a and b). For the TBM targets, out of the 8263 models, 5414 models have GDT-TS scores $\geq$70, while 884 are selected by SVM. Considering a GDC score $\geq$30 (this value is selected based on the distribution of GDC scores; data not shown) 4020 qualified as good models. Interestingly, 632 models are selected by all the methods [Venn diagram (Fig. 6a)], indicating that most of the models (71.4%) selected by SVM could be easily validated by the GDT-TS scores. In the case of FM, a GDT-TS score $\geq$30 is considered for selection and again about 86.6% (13 out of 15 selected by SVM) have a high GDT-TS score. Further, the network properties of the models predicted by each method and their combinations are examined. Fig. 6(c–f) plot the different network parameters as a function of I_{min} for models qualified by (a) all the methods [blue], (b) only SVM [green], (c) only GDC [cyan], (d) only GDT-TS [red] and (e) models not selected by any of the scoring scheme [black]. As expected, the characteristic patterns of native structures, described before in this work, are distinctively shown by the blue line [selected by all methods] while the black line [selected by none] deviates significantly from this pattern. Furthermore, models selected by any one of the methods display transition patterns with minor variations. Amongst these, models picked up by SVM show a pattern closer to the characteristic profile described earlier, followed by models selected by GDC and GDT-TS. The transition profile of GDC and SVM selected models seem to satisfy all the criteria of a characteristic profile, perhaps emphasising the importance of including side chain atoms for quality assessment.

To evaluate the quality of models picked by different methods, RMSD distribution of the models at the C_α and all-atom level is examined. Fig. 7(a and b) show the RMSD distribution of models selected by (left) only GDT-TS, (middle) both GDT-TS and SVM and (right) only SVM as a pie chart. At the C_α level (Fig. 7a), models selected by GDT-TS have RMSD mainly in the range of 1.5 Å–3.0 Å [only GDT-TS (55%) and common panel (69%)]. Models with RMSD >3 Å and $\leq$5 Å have also been significantly selected by the GDT-TS scoring scheme. For the SVM selected models, the majority fall in the range of 1.5 Å–3 Å (69% + 18%) and 3 Å–5 Å (28% + 65%), a considerable percentage (17%) of models also have a RMSD value below 1.5 Å. This indicates that, while the network based method selects only a small number of models as compared to GDT-TS, it judiciously selects models with high accuracy as seen by the high percentage of models selected in the low RMSD range. Similar behaviour is seen for RMSD-all atom distribution. Strikingly, in this case models selected only by SVM show the lowest RMSD value (15%), while GDT-TS fails to pick any model with such a low value of RMSD. However, SVM also picks up a significant number of models as compared to GDT-TS with RMSD-all > 5 Å. This is a consequence of the fact that the network based method performs independent of the native structure and is solely based on the interaction strength of the side chain atoms, proper packing and orientation. The large RMSD in SVM selected models may arise due to different side-chain conformations and/or loop orientations (also may be due to false positives), indicating the possibility of structures with alternate conformations. An example of this is shown in Fig. 7c, where the model (T0667TS028_2) selected only by SVM for target Id T0667 has C_α-RMSD value = 2.64 Å and GDT-TS score = 54.82. Although side chain interaction is the major feature that drives the present network method, backbone packing is also checked on the basis of main-chain hydrogen bonds. Fig. 3(b–d) represent the percentage distribution of MHB for models selected by GDT-TS, both methods and SVM. Most of the models selected by these methods have MHB in the preferred region, however in the case of GDT-TS; a significant proportion of these models also lie in the low probability zone. Another example presented is the target Id T0708 (4H17), for which all the models selected by SVM are also selected by the GDT-TS score. In this case, the selected models align extremely well with the native structure (Fig. S3, ESI†) and exhibit low RMSD values, in the range of 1.66 Å to 1.98 Å.

An interesting example is the target Id T0661 (4FCZ), in which the native structure is incomplete with N and C terminal residues missing. However several modellers have provided good models of the complete structure. Strikingly, SVM has selected these models [with RMSD range from 2.10 Å to 2.55 Å], whereas, GDT-TS has failed to identify these models. The native (blue) and the models (red) are superposed in Fig. 8. An impressive observation is that the models have an intact core very much similar to that of the native. The N and C terminal helices are on the periphery of this core in the models. It should be emphasised that the network method is able to pick up these models without a reference structure. This is mainly due to the conceptual framework that the native structures have a well-packed core and characteristic network profiles.

Software tool

The methodology described in this study has been implemented as an independent module (PSN_QA) of our PSN analysis web-server, GraProStr and is freely available at http://vishgraph.mbu.iisc.ernet.in/GraProStr/PSN-QA.html. A screenshot of the home page of the tool is shown in Fig. 9a. It takes a structure in pdb format as input and calculates the network properties at different I_{min}s. The parameters are then passed on to the SVM classifier for a quality check. The tool outputs the quality of the

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

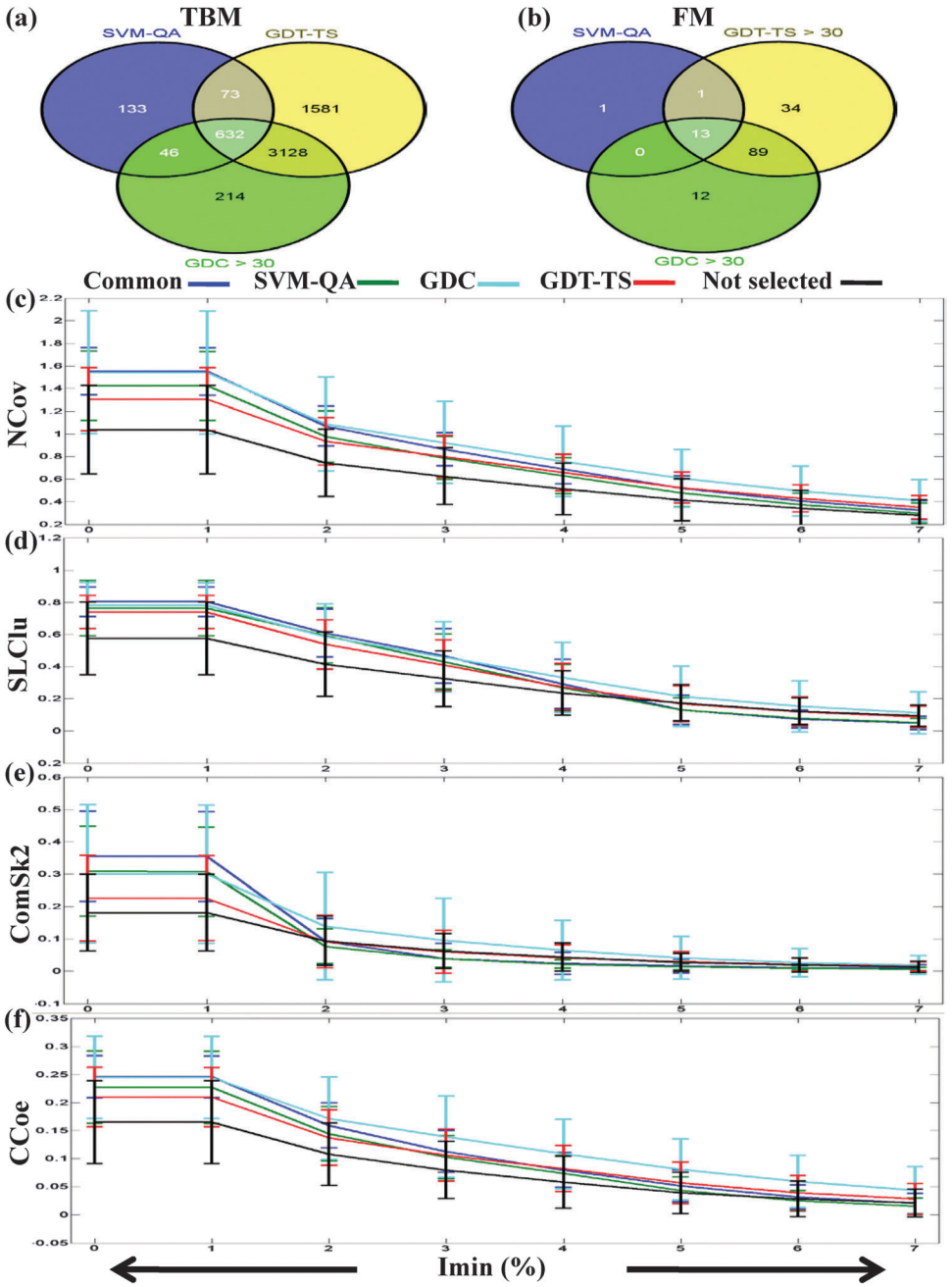

Fig. 6 SVM prediction for CASP10 dataset and comparison with CASP10 assessment. Venn diagrams representing the number of CASP10 models predicted by SVM, GDT-TS and GDC scores for (a) Template Based Modelling (TBM) and (b) Free Modelling (FM) targets. Network parameters (c) NCov, (d) SLClu, (e) ComSk2 and (f) CCoe are also plotted as a function of I_{min} for each category. *X* and *Y* axes are similar to that of Fig. 5.

input structure as native or non-native like (Fig. 9b). For structures with multiple chains, a quality check is provided for each subunit. Moreover, files containing the network parameter values at different I_{min}s and plots showing the transition profile are also available for download.

Discussion

Proteins acquire their unique structures through a subtle balance of various energetic and topological factors. The secondary structures such as helices and sheets, stabilized by the backbone hydrogen bonds, ensure the packing of the polypeptide chain. In addition, the unique structure is stabilized by non-covalent interactions of the amino acid side-chains, which should be optimal at the global level. PSN, being an all atom representation, captures not only the global topological features but also the chemistry of the side-chains. Our earlier study of PSN at low interaction strength[36] was the first work in this area, which indicated the potential of the network method to distinguish decoy structures from the native ones.

 This journal is © The Royal Society of Chemistry 2013

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

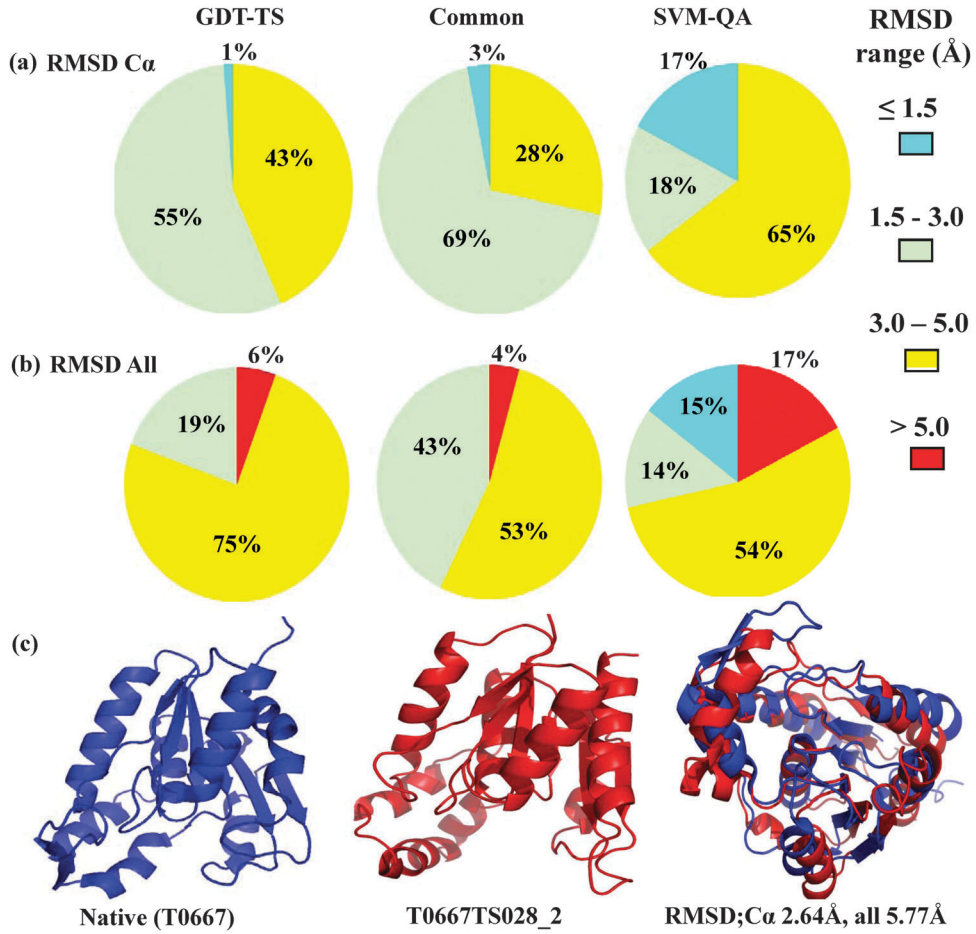

Fig. 7 Comparison of CASP10 models predicted as good quality by GDT-TS [left panel], both GDT-TS and SVM [middle panel] and only SVM [right panel]. Percentage distribution of the CASP10 models, selected by the two methods and their combinations, based on (a) C_α-RMSD and (b) RMSD-all. (c) An example structure of target Id T0667 (left panel) and model T0667TS028_2 (selected only by SVM) (middle panel) in which the secondary structures and their mutual orientations are well superposed (right panel). High RMSD arises due to the reorientation of the loops.

Also, by extensive data analysis, we had demonstrated that the transition profiles of the network parameters, such as the largest cluster or community, as a function of interaction strength (I_{min} = 0% to 7%), are the hallmarks of the native protein structures.[14] Investigation of protein structures incorporating this feature would discriminate the natives from decoys in a more accurate manner. Furthermore, it also lends support to the observation that the native structures indeed possess general topological features while they are absent in random models. And the present study has shown that the network profiles of decoys/models in which the side-chain packing is not optimal, deviate from those exhibited by native structures.

Specifically, in the present study, we have explored the side-chain network profiles for a large number of native proteins and decoy/modelled structures, and the backbone packing is considered through main-chain hydrogen bonds. The size of the largest community as a function of I_{min} represented in Fig. 10 for the native and the decoys/models clearly indicates the importance of the slope at the transition region and the relative values above and below the transition. The size of the community is larger at lower I_{min} [0–2%] and smaller at the higher I_{min} [>4%] in the native protein, when compared with the decoys, leading to a steeper slope for the native. The consistent display of the steep transition profile by the native structures is discussed in the results section by comparing with random models [Fig. S2, ESI†] which drastically deviate from the native, while the decoys/models exhibit behaviour close to the native structures. An interesting observation from the structural cartoons in Fig. 10 is that a greater number of residues involved in forming k-2 community lie on the secondary structures of the native protein, while they are largely positioned on the loops for decoy structures. The optimal balance between the side-chain interaction and the backbone packing is crucial for the uniqueness of the protein structure. These topological properties have been integrated in SVM to build a classifier and the importance of the contributing features is mathematically evaluated. The model has effectively captured the above mentioned features [accuracy = 94.22%, AUC = 0.9853], specifically identifying the transition related parameters as top contributors. The classifier has successfully identified the good-quality models from the pool of CASP10

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

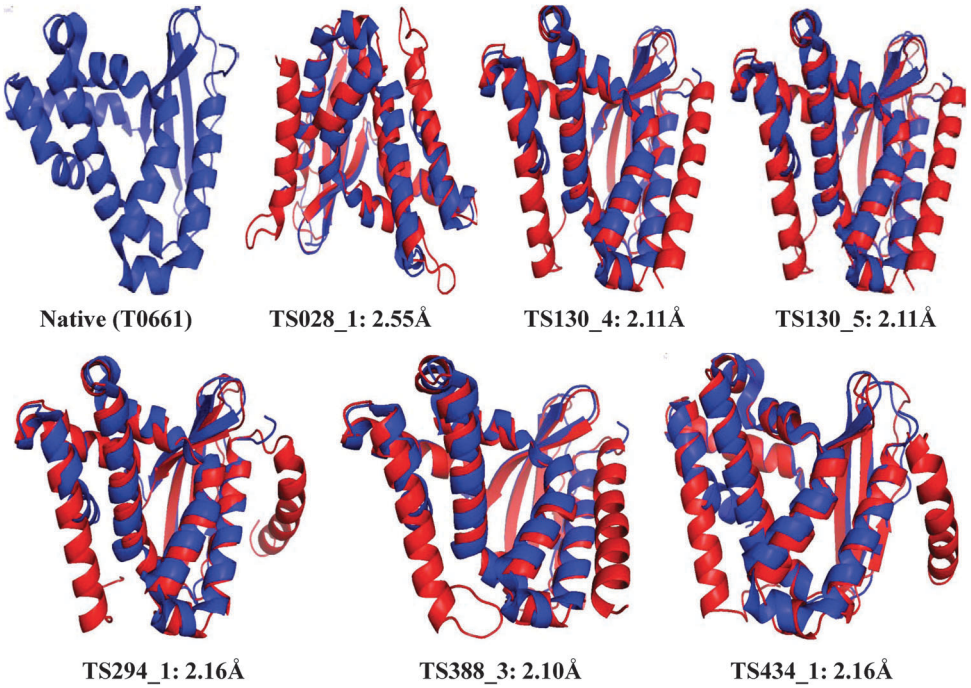

Fig. 8 Comparison of model structures predicted by SVM and the native structure for target Id T0661 (4FCZ). The native structure in these figures is shown in blue while the SVM predicted models are shown (in red) aligned with the native structure. The names of the modelled structures as provided in CASP10 are used. C_α-RMSD value for each structure is also provided.

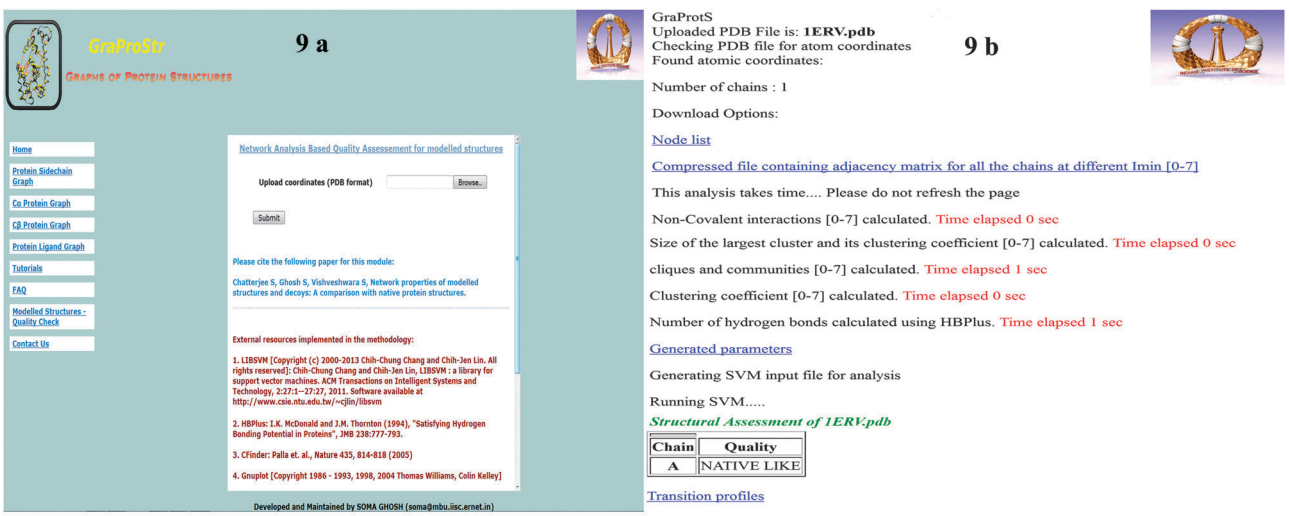

Fig. 9 Screenshot of the web server (http://vishgraph.mbu.iisc.ernet.in/GraProStr/PSN-QA.html), to check the quality of the model. Panel (a) shows the home page of the web server and panel (b) shows the output results.

predicted structures. This is the first rigorous attempt to make use of the global topological features unique to the native structures, to assess modelled structures. In addition, the present study reinforces the fact that the native proteins acquire their unique structures by side-chains interactions, leading to optimal global connections at different interaction strengths. This is captured by the evaluation of network parameters on the PSNs of a large number of native proteins and the decoys, by and large fail to exhibit the subtleties in the network profiles. The trained SVM model has successfully captured the subtle features.

It is to be pointed out that the methodology developed here can be easily used for quality assessment of the models predicted in the absence of native structures. The present work has shown that the native structures exhibit characteristic topological features [measured from several network parameters] such as (a) higher value at lower I_{min}, (b) lower value at higher I_{min} and most importantly (c) steep transition from

 This journal is © The Royal Society of Chemistry 2013

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

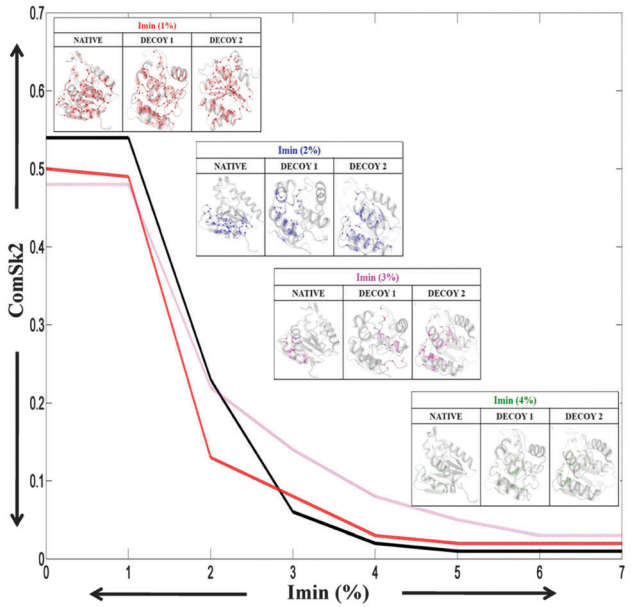

Fig. 10 Transition profile (I_{min} = 1%–4%) of an example native structure (4FLE) (black) and two randomly chosen modelled structures (red, pink) from the CASP10 dataset for ComSk2. The inset panels show the structures of the protein and the two decoys. Spheres in the structures represent the residues involved in forming the ComSk2 and a decrease is seen with transition from I_{min} = 1% to 4%. It is also noted that the residues in the case of the decoy structure lie dominantly on the loop region.

I_{min} = 0% to I_{min} = 4%. Here we have performed assessment of the CASP10 predicted models for 83 targets and the detailed comparison of SVM selected models with others are shown in Fig. 6–8 and Fig. S3 (ESI†). Overall, the classifier is able to pick up structures that show native-like patterns such as good packing, well-formed secondary structures and proper orientation of the secondary structures, reiterating the fact that analysing parameters at the backbone level, may not be sufficient and a rigorous analysis at the backbone as well as the sidechain level may be required to distinctively differentiate a native structure from the decoy/model structure. SVM is also able to pick up a larger percentage of models that are closer to the native and we believe that integration of the network model with the existing pipeline procedures will enhance the predictability of the already available tools.

Conclusion

To summarise, a large number of native structures and decoy/model structures were used to generate protein structure networks at different interaction strengths [I_{min}]. The profiles of the network properties as a function of I_{min} for native proteins and the decoy/model structures are compared to identify patterns specific to native structures. The network parameters generated at different I_{min} and main-chain hydrogen bonds are integrated into Support Vector Machine to develop a classifier. An accuracy of 94.11% is obtained. The feature selection methods have captured the main chain hydrogen bonds and transition profiles of network parameters as the top contributors to the native structures. Classifier is further validated with specific test cases and later used for predicting the quality of CASP10 predicted models. Thus, a robust and general classifier to distinguish native protein from decoy/model is built using PSNs at different interaction strengths, which can be used for building good models and also to assess predicted models. A web-server (http://vishgraph.mbu.iisc.ernet.in/GraProStr/PSN-QA.html) has been made available for this purpose. The study also reinforces the observation that the native structures have unique global-topological transition profiles.

Acknowledgements

We acknowledge the computational resources provided the Department of Biotechnology (DBT), through the Centre of Excellence (COE). SC thanks Department of Science and Technology (Mathematical Biology Grant, IMI) for fellowship. SV acknowledges the Council of Scientific and Industrial Research (CSIR, India) for Emeritus Scientist, CSIR.

References

1 C. B. Anfinsen, *Science*, 1973, **181**, 223–230.
2 P. Lewis, D. Kotelchuck and H. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 1970, **65**, 810–815.
3 S. Burley and G. Petsko, *Science*, 1985, **229**, 23.
4 H. J. Dyson, P. E. Wright and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 13057–13061.
5 J. T. Kellis and D. Kerstin Nyberg, *Nature*, 1988, **333**, 784–786.
6 C. N. Pace, H. Fu, K. L. Fryar, J. Landua, S. R. Trevino, B. A. Shirley, M. M. N. Hendricks, S. Iimura, K. Gajiwala and J. M. Scholtz, *J. Mol. Biol.*, 2011, **408**, 514–528.
7 L. Serrano, M. Bycroft and A. R. Fersht, *J. Mol. Biol.*, 1991, **218**, 465–475.
8 G. D. Rose, P. J. Fleming, J. R. Banavar and A. Maritan, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 16623–16633.
9 K. A. Dill, S. B. Ozkan, M. S. Shell and T. R. Weikl, *Annu. Rev. Biophys.*, 2008, **37**, 289.
10 C. M. Dobson, *Nature*, 2003, **426**, 884–890.
11 A. R. Fersht, *Nat. Rev. Mol. Cell Biol.*, 2008, **9**, 650–654.
12 M. Karplus, *Nat. Chem. Biol.*, 2011, **7**, 401–404.
13 C. Atilgan, O. B. Okan and A. R. Atilgan, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 3363–3375.
14 G. Bagler and S. Sinha, *Phys. A*, 2005, **346**, 27–33.
15 A. R. Atilgan, P. Akan and C. Baysal, *Biophys. J.*, 2004, **86**, 85.
16 L. H. Greene and V. A. Higman, *J. Mol. Biol.*, 2003, **334**, 781–791.
17 M. Vassura, L. Margara, P. Fariselli and R. Casadio, *Artif. Intell. Med.*, 2009, **45**, 229–237.
18 T. J. Taylor and I. I. Vaisman, *Phys. Rev. E*, 2006, **73**, 041925.
19 A. Küçükural, U. Sezerman and A. Erçil, *Adv. Bioinf. Comput. Biol.*, 2008, **6**, 59–67.
20 N. Kannan and S. Vishveshwara, *J. Mol. Biol.*, 1999, **292**, 441–464.

Published on 13 May 2013. Downloaded by Swinburne University of Technology on 29/04/2017 05:33:34.

21 A. Sukhwal, M. Bhattacharyya and S. Vishveshwara, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2011, **67**, 429–439.
22 R. Sathyapriya, M. Vijayabaskar and S. Vishveshwara, *PLoS Comput. Biol.*, 2008, **4**, e1000170.
23 D. Deb and S. Vishveshwara, *Biophys. J.*, 2009, **97**, 1787–1794.
24 S. Wu, J. Skolnick and Y. Zhang, *BMC Biol.*, 2007, **5**, 17.
25 J. Zhang and Y. Zhang, *PLoS One*, 2010, **5**, e15386.
26 E. S. Huang, S. Subbiah, J. Tsai and M. Levitt, *J. Mol. Biol.*, 1996, **257**, 716–725.
27 B. Park and M. Levitt, *J. Mol. Biol.*, 1996, **258**, 367–392.
28 B. Ranjit and C. Pinak, *BMC Struct. Biol.*, 2009, **9**, 76–84.
29 D. Gilis, *J. Biomol. Struct. Dyn.*, 2004, **21**, 725–735.
30 R. Zhou, B. D. Silverman, A. K. Royyuru and P. Athma, *Proteins*, 2003, **52**, 561–572.
31 M. Kalman and N. Ben-Tal, *Bioinformatics*, 2010, **26**, 1299–1307.
32 R. P. Bahadur and P. Chakrabarti, *BMC Struct. Biol.*, 2009, **9**, 76.
33 A. Nath Jha, S. Vishveshwara and J. R. Banavar, *Protein Sci.*, 2010, **19**, 603–616.
34 S. Miyazawa and R. L. Jernigan, *Proteins: Struct., Funct., Bioinf.*, 1999, **34**, 49–68.
35 J. Moult, K. Fidelis, A. Kryshtafovych and A. Tramontano, *Proteins: Struct., Funct., Bioinf.*, 2011, **79**(Suppl 10), 1–5.
36 S. Chatterjee, M. Bhattacharyya and S. Vishveshwara, *J. Biomol. Struct. Dyn.*, 2012, **29**, 1110–1126.
37 K. Brinda and S. Vishveshwara, *Biophys. J.*, 2005, **89**, 4159.
38 W. S. Noble, *Nat. Biotechnol.*, 2006, **24**, 1565–1567.
39 J. Guo, H. Chen, Z. Sun and Y. Lin, *Proteins: Struct., Funct., Bioinf.*, 2004, **54**, 738–743.
40 S. Hua and Z. Sun, *J. Mol. Biol.*, 2001, **308**, 397–408.
41 J. Ward, L. J. McGuffin, B. F. Buxton and D. T. Jones, *Bioinformatics*, 2003, **19**, 1650–1655.
42 H. Kim and H. Park, *Protein Eng.*, 2003, **16**, 553–560.
43 Y. D. Cai, X. J. Liu, X. Xu and K. C. Chou, *Comput. Chem.*, 2002, **26**, 293–296.
44 J. R. Bradford and D. R. Westhead, *Bioinformatics*, 2005, **21**, 1487–1494.
45 Q. Dong, Y. Chen and S. Zhou, *International Journal of General Systems*, 2011, **40**, 417–425.
46 P. Mereghetti, M. L. Ganadu, E. Papaleo, P. Fantucci and L. De Gioia, *BMC Bioinf.*, 2008, **9**, 66.
47 R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. M. Strauss and D. Baker, *Proteins: Struct., Funct., Bioinf.*, 2002, **45**, 119–126.
48 J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl and D. Baker, *Proteins: Struct., Funct., Bioinf.*, 2003, **53**, 76–87.
49 R. Samudrala and M. Levitt, *Protein Sci.*, 2008, **9**, 1399–1401.
50 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
51 T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introduction to algorithms*, MIT press, 2001.
52 M. Bhattacharyya, A. Ghosh, P. Hansia and S. Vishveshwara, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 506–517.
53 G. Palla, I. Derényi, I. Farkas and T. Vicsek, *Nature*, 2005, **435**, 814–818.
54 C. C. Chang and C. J. Lin, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011, **2**, 27.
55 I. K. McDonald and J. M. Thornton, *J. Mol. Biol.*, 1994, **238**, 777–793.
56 Y. W. Chen and C. J. Lin, *Feature Extraction*, 2006, 315–324.
57 M. Robnik-Šikonja and I. Kononenko, *Mach. Learn.*, 2003, **53**, 23–69.
58 V. Mariani, F. Kiefer, T. Schmidt, J. Haas and T. Schwede, *Proteins: Struct., Funct., Bioinf.*, 2011, **79**(Suppl 10), 37–58.
59 L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao and N. V. Grishin, *Proteins: Struct., Funct., Bioinf.*, 2011, **79**(Suppl 10), 59–73.
60 B. Monastyrskyy, K. Fidelis, A. Tramontano and A. Kryshtafovych, *Proteins: Struct., Funct., Bioinf.*, 2011, **79**, 119–125.
61 A. Kryshtafovych, K. Fidelis and A. Tramontano, *Proteins: Struct., Funct., Bioinf.*, 2011, **79**(Suppl 10), 91–106.
62 V. Mariani, F. Kiefer, T. Schmidt, J. Haas and T. Schwede, *Proteins*, 2011, **79**(Suppl 10), 37–58.
63 S. N. Soffer and A. Vázquez, *Phys. Rev. E*, 2005, **71**, 057101.

 This journal is © The Royal Society of Chemistry 2013