

Structural bioinformatics

Prediction of viable circular permutants using a graph theoretic approach

Konrad H. Paszkiewicz^{1,*}, Michael J. E. Sternberg¹ and Michael Lappe²¹Structural Bioinformatics Group, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, UK and ²Max-Planck-Institute for Molecular Genetics, Illhnestrasse 63-73, 14195 Berlin, Germany

Received on January 11, 2006; revised and accepted on March 9, 2006

Advance Access publication March 16, 2006

Associate Editor: Keith A Crandall

ABSTRACT

Motivation: In recent years graph-theoretic descriptions have been applied to aid the analysis of a number of complex biological systems. However, such an approach has only just begun to be applied to examine protein structures and the network of interactions between residues with promising results. Here we examine whether a graph measure known as closeness is capable of predicting regions where a protein can be split to form a viable circular permutant. Circular permutants are a powerful experimental tool to probe folding mechanisms and more recently have been used to design split enzyme reporter proteins.

Results: We test our method on an extensive set of experiments carried out on dihydrofolate reductase in which circular permutants were constructed for every amino acid position in the sequence, together with partial data from studies on other proteins. Results show that closeness is capable of correctly identifying significantly more residues which are suitable for circular permutation than solvent accessibility. This has potential implications for the design of successful split enzymes having particular importance for the development of protein–protein interaction screening methods and offers new perspectives on protein folding. More generally, the method illustrates the success with which graph-theoretic measures encapsulate the variety of long and short range interactions between residues during the folding process.

Contact: konrad.paszkiwicz@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The increase in the amount and complexity of biological data in recent years has resulted in a landscape that is well-suited for analysis via graph-theory methods. These have often been used to help with the study and planning of high-throughput proteomics experiments (Bradshaw and Burlingame, 2005; Betton, 2004; Gilbert, *et al.*, 2004; Lappe and Holm, 2004) and have themselves generated large datasets well-suited for a graph-based representation (Xenarios *et al.*, 2000). These networks have been analyzed using such an approach where the nodes represent individual proteins/domains and whilst the edges represent interactions between them (Borgwardt *et al.*, 2005; Santonico *et al.*, 2005). However the study of individual protein structures as residue interaction graphs (RIGs), where residues are defined as nodes and the interactions

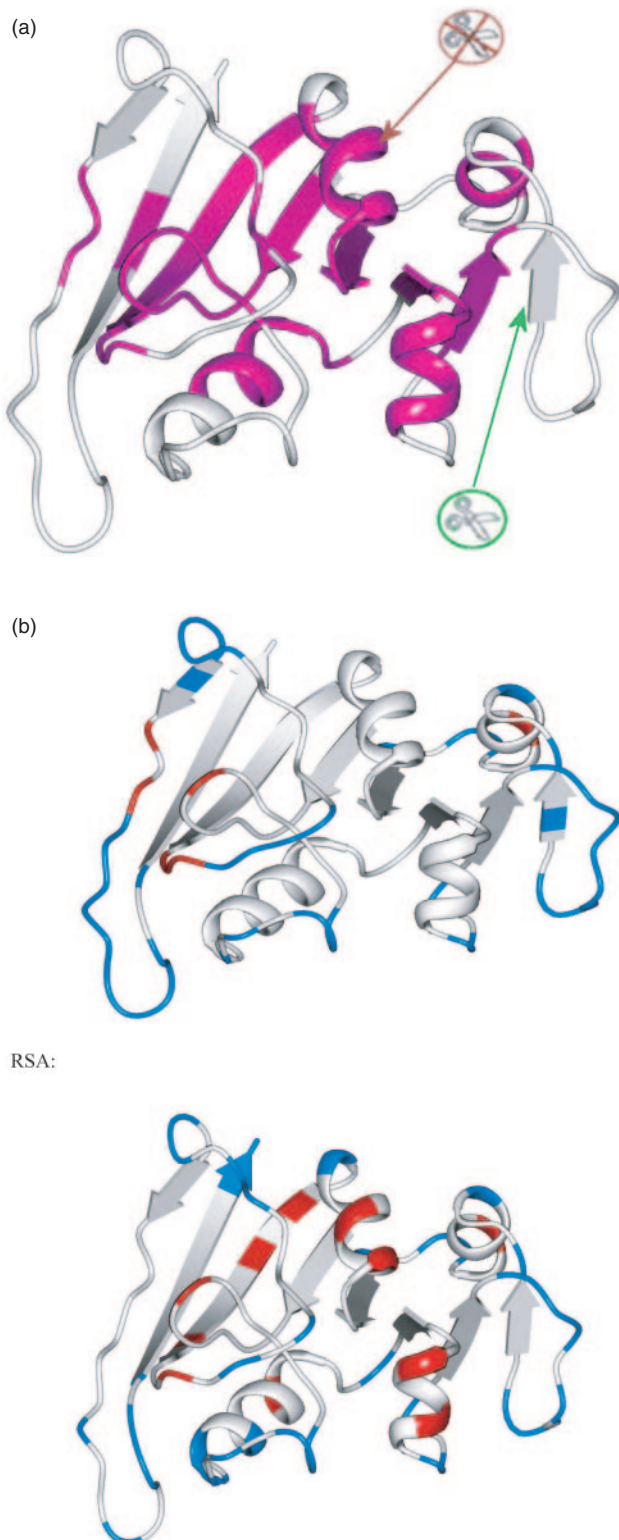
between them as edges, are a more recent development. A number of important results have been published highlighting the usefulness of this approach to enhance our understanding and modeling of sequence–structure–function relationships. Vendruscolo *et al.* (2001, 2002) demonstrated using graph-methods together with molecular dynamic simulation that highly central residues act as nucleation centers in the protein. More recently Amitai *et al.* (2004) showed that a measure known as closeness (see below) correlates well with the position of active sites in proteins. Del Sol *et al.* (2005) have identified clusters of highly central residues at or near the protein–protein interfaces. Both these results reflect the fact that proteins act in a concerted fashion, with conformational change being transmitted via a few key residues which are central to the network of residue interactions.

Here we investigate the usefulness of graph-theory to describe the relationship between protein sequence, structure and function by constructing graphs of protein structures (see Methods). We then apply the closeness graph-measure to predict the location of suitable split sites which result in viable circular permutants. Circular permutation involves the covalent linkage of the N and C termini of a protein and cleavage of the backbone at a given position to produce new termini (Fersht, 1999). This has particular relevance for the study of sub-domain regions termed folding elements (FEs). These were first discovered by Iwakura *et al.* (2000) using the globular protein dihydrofolate reductase (DHFR) from *Escherichia coli*. These regions are intolerant to disruption via circular permutation, with the result that they do not fold to a functional native-like protein. In DHFR a total of 10 contiguous blocks of 2–14 residues are formed which do not directly correspond to either buried regions or secondary structure elements (Fig. 1a). The precise significance of these regions is as yet unknown, although a recent study has proposed a folding mechanism based around these elements (Arai *et al.*, 2003). There is additional support for this hypothesis in that many early folding residues (EFRs) are found within the FEs (Arai *et al.*, 2003; Jones and Matthews, 1995). Unfortunately because the experimental effort involved in characterizing these FEs is considerable, it is so far the only protein for which a complete set of FEs may be assigned.

Using the wild-type native structure of DHFR with the above dataset, together with partial data available from other circular permutation studies on other proteins, we are able to show that closeness is able to predict the location of suitable split sites more reliably than either burial or sequence conservation. More intriguingly, it also offers future potential to predict the precise

*To whom correspondence should be addressed.

boundaries of FE and non-FE regions (respectively where circular permutation does and does not lead to a functional protein) and thus avoid the need for extensive experimental studies of large numbers of other proteins.



This is of particular practical importance for the design of split reporter proteins and the development high-throughput protein interaction detection systems. In such systems a reporter protein (e.g. with fluorescent read-out) is split in two fragments and each is fused to one of the two proteins under study. If the two proteins do interact, the reporter fragments are brought together and refold. The reconstituted activity is used as a read-out for the protein-interaction. The design of such split-enzymes is not straightforward, and several splits need to be tested to find a viable split site (Galarneau *et al.*, 2002).

2 METHODOLOGY

The method represents a protein structure as a RIG in which residues are considered to be nodes and the interactions between them are defined as edges. Shortest paths between all pairs of residues can then be calculated (see below), and a value for the closeness of each residue obtained. Closeness is a measure of the average number of edges which must be traversed in order to reach any other node. It is noteworthy that we tested other measures similar to closeness (e.g. betweenness) and obtained similar results (data not shown) with closeness showing the best performance. Residues that are central to the network will have a shorter average shortest path length to other residues. DHFR was shown to have 73 FE residues out of a total of 159 residues (Iwakura *et al.*, 2000).

Graph generation

Co-ordinates were taken from the Protein Databank (PDB) for use in our analysis (PDB codes given in Supplementary information) (Berman *et al.*, 2002). RIGs were constructed with nodes representing individual residues while an edge represents an interaction between two residues. An interaction between residues was defined to occur if any heavy atom in residue i was within 4 \AA of any atom in residue j . Larger thresholds were also tested but had no qualitative impact upon the results (data not shown). Unweighted and undirected graphs were used. Where proteins consisted of multiple domains, the domain boundaries were calculated using the 'domain' program (Sternberg *et al.*, 1995) and separate graphs were generated for each domain. In the case of DHFR the SCOP database (Murzin *et al.*, 1995) identifies the protein as having a single domain. However, we considered it to be a bilobal protein with two domains (residue numbers 1–127, 128–159) and best results for relative side-chain area (RSA) and closeness were obtained by calculating separate graphs of the domains. Without this domain separation, the surface exposed FE at position V136-S138 in DHFR is not detected by RSA or closeness (data not shown).

Measures and parameters

Shortest paths (as per Fig. 2) were determined between all pairs of residues in each structure using Dijkstra's algorithm (Dijkstra, 1959). These paths were then used to calculate closeness values for each residue in the protein. Closeness is defined as the average length of all the shortest paths emanating

Fig. 1. Predicting non-FEs in DHFR. (a) FE positions in DHFR. Three-dimensional cartoon representations of DHFR showing the positions of the FEs (magenta). The scissors indicate that the FE regions do not tolerate disruption through circular permutation. All of the regions in which FEs exist are covered by the closeness measure. This suggests that closeness could be used to predict non-FE regions which reliably result in viable circular permutants. [Images courtesy of Suhail Islam using Prepi (www.sbg.bio.ic.ac.uk)]. (b) Top 50 predictions of non-FE positions in DHFR. Regions in blue are regions correctly predicted to yield non-FEs. Those in red indicate sites that were predicted to be non-FEs but were experimentally shown to be part of FEs. 11 predictions are incorrect in the top 50 when using RSA whilst only 5 are incorrect when using closeness.

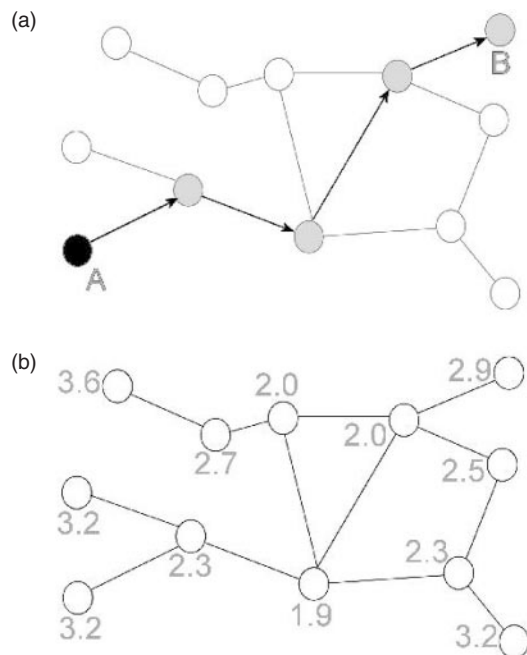


Fig. 2. Describing closeness. (a) To calculate the closeness of node A within the interaction network we first calculate the shortest path lengths to all the other nodes in the network. For clarity only the path A→B (length 4) is shown. One would then need to continue to find the shortest paths between A and the remaining nodes in the network. Once all the shortest path lengths for A have been calculated we perform the same procedure using all the other nodes as starting points. (b) The value of closeness for a node is assigned by summing the shortest path lengths and then dividing by the number of paths.

from a particular node to all other nodes in the graph. Buried residues will thus tend to have lower raw closeness values than those at the surface of the protein, with the exception of functional residues (Amitai, 2004 #10). RSA values were calculated using the program Naccess using default parameters (Hubbard *et al.*, 1991). Z-scores were derived using the raw scores output from each of the three measures tested. For a set of values X with standard deviation σ and average μ , a Z-score for a value x was defined as $(x - \mu) / \sigma$. Z-scores were obtained such that a Z-score > 0 for a given residue indicated that the residue was a member of a FE.

Statistical methods

In prediction those residues with Z-scores > 0 were considered to be FE regions whilst those with Z-scores ≤ 0 were not. Point biserial correlation coefficients (Edwards, 1976) were calculated using the *R* statistical package (Ihaka, 1996). Rather than calculating the correlation between two continuous variables, the point biserial coefficient calculates the correlation between a binary variable (here FE or non-FE) and a continuous variable (e.g. closeness, RSA or sequence conservation). The values of point biserial coefficients are similar to linear least-square correlation coefficients.

Sequence conservation

Sequence conservation was quantified by the calculation of sequence entropies for each equivalent residue in an alignment. Relative sequence entropies were calculated using a position specific scoring matrix derived from three iterations of PSI-BLAST (Altschul *et al.*, 1997).

Entropies for each residue position were then calculated using

$$H = - \sum_{i=1}^{20} P_i \log_2 P_i,$$

where P_i denotes the probability that a given amino acid type appears at a given position.

$$P_i = \alpha_i \exp^{\lambda x_i}.$$

The subscript i corresponds to each of the 20 amino acids, α_i to the background frequency of the amino acid over the entire dataset, x_i to the residue type frequency at the position and λ is a scaling factor equal to 0.318 derived from the BLOSUM62 matrix (Henikoff and Henikoff, 1992).

Receiver operating characteristic curves

Receiver operating characteristic (ROC) curves were calculated for sequence entropy, RSA and closeness in DHFR to test their power to predict the status of a given residue (FE or non-FE). True positives were defined as experimentally determined non-FE residues with a Z-score for a given measure ≤ 0 or FE residues with a score > 0 . The area under curve (AUC) for each of the measures was calculated using a trapezoidal-based rule. To calculate whether differences between curves were statistically significant, standard errors for the AUC were calculated using the method described by Hanley and McNeil (1982). A score for a given pair of AUCs is calculated using the following:

$$w = \frac{A1 - A2}{SE(A1 - A2)},$$

where $A1$ and $A2$ are the AUC for the respective curves, $SE(A1 - A2)$ is the standard error of the difference between the areas $A1$ and $A2$ as described by Hanley and McNeil in further work (Hanley and McNeil, 1982, 1983). Scores > 1.96 are significantly different at the 5% level.

Experimental data

Permutant data are shown in Supplementary information, Table 1. EFRs were identified by Jones and Matthews (1995) and were defined to be residues which were protected from hydrogen exchange within 13 ms of folding commencing.

FES in DHFR (PDB id: 1RX4) were located by Iwakura *et al.* (2000) at the following positions: Ser 3-Ile 14, Trp 30-Leu 36, Val 40-Ser 49, Arg 57-Ser 63, Glu 80-Ala 83, Ile 91-Leu 104, Ala 107-Glu 118, His 124-Phe 125, Val 136-Ser 138 and Glu 154-Ile 155.

Functional permutants in DsbA (PDB id: 1A2J) were constructed by Hennecke *et al.* (1999) at positions Gly 65-Gly 66, Lys 118-Gly 119, Asn 127-Ser 128 and Gly 157-Asn 156. Non-functional permutants were constructed at positions Leu 72-Thr 73, Trp 76-Ala 77, Leu 82-Gly 83, Val 88-Thr 89, Glu 94-Gly 95, Tyr 122-Asp 123, Lys 132-Ser 133, Val 135-Ala 136, Ala 142-Ala 143 and Pro 151-Ala 152.

Functional permutants in GFP (PDB id 1EMB) were constructed by Topell *et al.* (1999) at positions Gly 24-His 25, Tyr 38-Trp39, Tyr 49-Tyr 50, Asp 102-Asp103, Gly 116-Asp 117, Asn 144-Tyr 145, Gln 157-Lys 158, Asp 173-Gly 174 and Gly 228-Ile 229. Non-functional permutants were constructed at positions Tyr 9-Gly 10, Leu 44-Lys 45, Tyr 63-Phe 64, Val 68-Gln 69, Phe 75-Asn 76, Asp 82-Phe 83, Arg 96-Tyr 97, Phe 130-Lys 121, Phe 165-Lys 166, Gln 204-Ser 205 and Asn 212-Glu 213.

Additional permutant data from other proteins is shown in Supplementary information, Table 1.

3 RESULTS

FE and EFR correlation with measures in DHFR

Point biserial correlations between the positions of the FEs and the Z-scores of closeness, RSA and sequence conservation were calculated and yielded values of 0.58, 0.39 and 0.21 respectively. Using a windowed average with a width of 5, the correlation of sequence conservation and RSA improve from 0.21 to 0.4 and 0.39 to 0.42 whilst the correlation for closeness decreased to 0.57. The difference between the higher correlation for closeness compared

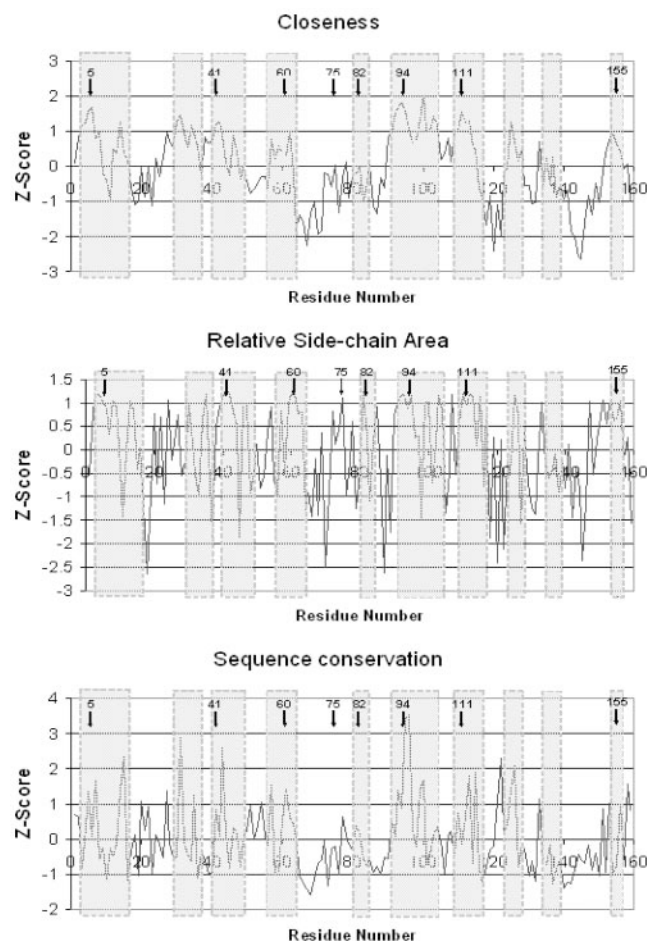


Fig. 3. Profiles of Closeness, RSA and Sequence conservation in DHFR. Closeness profiles calculated for individual domains of DHFR. The shaded regions represent FE regions. Arrows indicate the positions of experimentally detected EFRs. Note that regions with high Z-scores that are close to the rest of the network tend to fall within FEs. RSA profiles calculated for individual domains of DHFR. As with closeness, regions containing FEs tend to have above average RSA Z-scores (i.e. which are buried) to the rest of the network. However the profile contains far more noise than closeness owing to the reduced correlation between neighboring residues. Sequence conservation contains much less information relating to the position of either FEs or EFRs.

with RSA is significantly different at 1% level. The correlation with the position of FEs for sequence conservation is statistically significant only when an average window is used. Figure 3 shows the profiles for closeness, RSA and sequence conservation scores for each residue position in DHFR, together with the positions of experimentally identified FEs and EFRs. Within each FE there is at least one major peak in the closeness profile. From the RSA profile one can also observe that these regions tend to be highly buried, but there are many more pronounced peaks in the RSA profile which do not correspond to EFRs or FEs. The sequence conservation profile is less capable of revealing the presence of EFRs or FEs. The only exception to this is Val 75 which has a pronounced peak in the RSA profile, but not the closeness profile. However this EFR was found not to be a part of a FE by Iwakura *et al.* (2000).

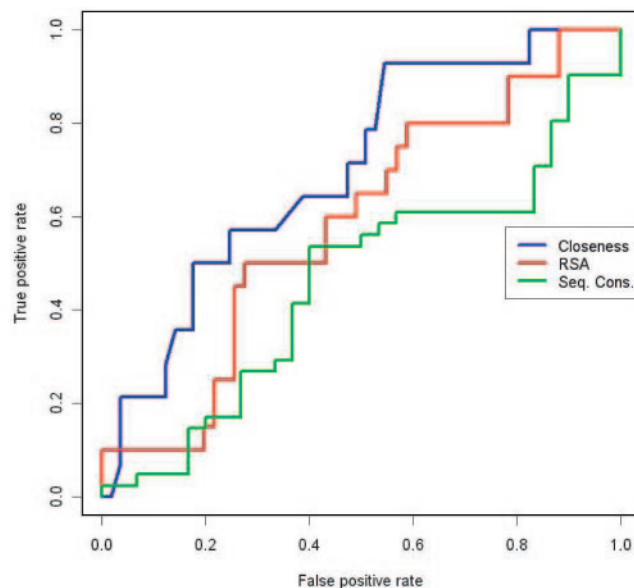


Fig. 4. Predicting non-FEs in DHFR. The predictive power of the closeness measure is above RSA and sequence entropy in the prediction of viable DHFR permutants (i.e. non-FEs). Shown is the ROC curve when each measure is used to predict whether a split at a particular residue will result in folded protein. The AUCs are 0.7, 0.58 and 0.49 for closeness, RSA and sequence conservation respectively. The difference between the curves is significant to the 1% level.

Owing to the limited experimental data on DHFR it is not possible to perform any further statistics on the relative success of closeness and RSA in the prediction of EFRs (26 of the 159 residues in DHFR were experimentally tested for early folding). However, the complete dataset for DHFR FEs allows us to investigate the application of closeness and RSA to predict the location of split sites for viable circular permutants in DHFR.

Predicting viable circular permutants

Since closeness and RSA correlate well with the presence of FEs in DHFR we now compare how successful each measure is in the prediction of viable split sites—i.e. the non-FEs. For the purposes of prediction with a given measure, non-FE residues are defined as having a Z-score ≤ 0 and FE residues having Z-scores > 0 . True positive results were defined as those where the prediction and the experimental observation were in agreement. Measures were ordered by ascending Z-score. Together with the experimental data on the location of FE and non-FE residues, ROC curves were plotted for each measure and the AUC for each curve calculated (Fig. 4). The results were 0.70 for closeness, 0.58 for RSA and 0.42 for sequence conservation. The difference between these AUCs was calculated as being significant to the 1% level between closeness, RSA and sequence conservation. For the first five predictions, closeness and RSA predict with 100% accuracy the location of viable split sites. However, none of the measures predict the same set of residues within this top five. Beyond this limit the closeness graph measure outperforms all other measures. Among the top 50 predictions RSA incorrectly predicts 11 residues as being viable split sites whilst closeness only predicts 5 incorrectly (illustrated in Fig. 1b).

Table 1. Additional permutant data

Protein	Split site	Closeness Z-score	RSA Z-score
Alpha spectrin SH3 (1UUE)	Ser 19	-2.00	+0.53
''	Asn 47	-0.40	-1.85
A-1-antitrypsin (1HP7)	Lys 368	-0.27	+0.12
SH6 (1RIS)	Asn 13	-1.02	-1.39
''	Ala 53	-1.48	-1.17
''	Pro 68	-1.15	+0.55
Chymotrypsin inhibitor (2CI2)	Pro 25	-0.28	-0.33
''	Leu 40	-0.10	+0.01
''	Val 53	-0.32	-0.30
Beta-lactamase (1TEM)	Glu 197	-0.84	-1.64

Measures of closeness and RSA for several experimentally verified viable circular permutants. Shown in brackets are the PDB codes used for graph construction. In all cases the experimental split sites used in the creation of permutants have closeness Z-scores <0 indicating non-FEs. However, using RSA in 4 out of the 10 cases the RSA measure has Z-scores >0 and incorrectly predicts permutants as being a FE.

Two other studies of circular permutants have yielded partial data and we use them to determine the generality of the results from DHFR. Partial studies on circular permutants have been performed on for disulfide oxidoreductase from *E.coli* (DsbA) (Hennecke *et al.*, 1999) and green fluorescent protein (GFP) (Topell *et al.*, 1999) from *Aqueorea Victoria*. In these experiments only ~10% of the possible permutants were produced and tested for their ability to fold. In summary, the available data indicates that closeness outperforms RSA in predicting viable circular permutants and that these results can be generalized beyond the case of DHFR. For GFP, we found that both closeness and RSA both correctly predicted all the eight permutants which did not fold. However, for DsbA, closeness correctly predicted 10 of 13 permutants whilst RSA predicted only 5 correctly. There is also added data from the design of successful circular permutants to study the folding of SH3, S6, CI2, antitrypsin and beta lactamase (Cobos *et al.*, 2003; Galarneau *et al.*, 2002; Grantcharova and Baker, 2001; Lee *et al.*, 2001; Lindberg *et al.*, 2002; Otzen and Fersht, 1998; Weikl and Dill, 2003) which amount in total to 10 circular permutants. Whilst all are correctly predicted by closeness, only a little over half are correctly predicted by RSA (Table 1).

As a final analysis to compare the predictive power of closeness and RSA we pooled all available circular permutant data to produce a dataset with a total of 188 split sites and observed which method predicted the correct result when the two measures disagreed. Closeness and RSA disagreed in their predictions in 49 cases. 29 of these were in the prediction of viable split sites. Of these, closeness was correct for a total of 21 times and whilst RSA was correct only 8 times in their predictions for these 29 cases. Thus closeness produces a greater number of correct predictions than RSA. To test whether this is statistically significant we use a binomial test, assuming that the probability that closeness makes a more accurate prediction is 0.5 if the two measures are indistinguishable. This yields a *P*-value of 0.01. Thus we conclude that closeness does produce a significantly greater number of correct predictions over RSA.

These results demonstrate that a graph-theoretic measure such as closeness encapsulates the regions essential to guide the folding

pathway to its native state by accounting for the network of interactions between residues throughout the protein as a whole.

4 DISCUSSION

We have shown that closeness can be used successfully to predict the location of viable split-sites (i.e. suitable locations for circular permutation) in DHFR and other proteins with greater success than either RSA or sequence conservation. The results using windowed averages demonstrate that the improvement is aided by the correlation between residues and the local neighborhood of interactions which are formed. If FEs are present in proteins generally it will be possible to avoid EFRs and the FEs which surround them by selecting residues with low centrality as determined by the closeness measure. Note that this is accomplished purely by measurements taken from the network of interactions between residues. The peaks in the closeness profile within each FE coincide with the position of EFRs.

Apart from the superior performance in predicting split sites, the clear peaks in the closeness profile in Figure 3 at the positions of EFRs indicate this method will simplify the identification of potential EFRs. Sequence conservation was shown not to identify the location of EFRs or FEs in the majority of cases.

The closeness measure is clearly detecting the increased number of interactions in the core and the susceptibility of the folding to the disruption of hydrophobic interactions and their exposure to solvent. These regions make up the immutable FEs which flank EFRs. The enhanced accuracy provided by closeness highlights the usefulness of a graph-based approach to protein structure analysis over and above physical measures such as RSA and sequence conservation. While an experienced practitioner might easily predict a few viable split-sites by examining the three-dimensional structure and selecting exposed loops, our method is capable of predicting more obscure non-FE regions within secondary structures, although these are often lower down the list of suggested sites. There is still room for improvement for our method with respect to the prediction of the precise boundaries of FEs. Further research here will contribute to the understanding of folding mechanisms.

Figure 3 also highlights the fact that EFRs occur in regions with enhanced closeness. Many EFRs have also been classified as binding sites in DHFR for a variety of small molecules. Of the 10 FEs in DHFR, 7 contain known small-molecule binding sites (Kinoshita and Nakamura, 2004) and the remaining 3 are experimentally classified as EFRs (Jones and Matthews, 1995). This finding suggests the existing hypothesis that individual FE-like subunits, each with specific binding capability, evolved independently in evolutionary history and were subsequently integrated into an ever larger and more complex functional unit. Such a relationship could be encoded in the network of interactions which surround each EFR. Indeed the work by Amitai *et al.* (2004) has shown that closeness can be used in conjunction with solvent accessibility to identify functional residues. If selection acts on this network of interactions we may gain new insight by developing a representation of these interactions which can then be compared tractably with other protein structures. If these interactions within FEs can be shown to be more highly conserved than other interactions in the protein then we may obtain evidence for this process.

The method described here demonstrates the usefulness of a graph-based approach to represent protein structures and that

graph-measures can reflect important physical properties detectable experimentally. The method is also of use to experimentalists who require additional evidence that a given site for circular permutation will prove viable. The importance of FEs in the folding process is highlighted by the enhanced closeness in the region of EFRs. The interactions formed between residues surrounding these EFRs are crucial to direct the folding pathway by generating suitable local environments early in the folding process. The resulting network of interactions between residues reflects these environments—any disruption to certain subnets cannot be tolerated if the protein is to fold to the native state. More generally however, the method has illustrated the use of graph-theoretic measures. Of particular importance is the ability of closeness and similar measures to encapsulate both local and global features of protein structures via residue–residue interactions.

ACKNOWLEDGEMENTS

K.P. is supported by the Medical Research Council.

Conflict of Interest: none declared.

REFERENCES

- Altschul,S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Amitai,G. et al. (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, **344**, 1135–1146.
- Arai,M. et al. (2003) Testing the relationship between foldability and the early folding events of dihydrofolate reductase from *Escherichia coli*. *J. Mol. Biol.*, **328**, 273–288.
- Berman,H.M. et al. (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
- Betton,J.M. (2004) High throughput cloning and expression strategies for protein production. *Biochimie.*, **86**, 601–605.
- Borgwardt,K.M. et al. (2005) Protein function prediction via graph kernels. *Bioinformatics*, **21**, i47–i56.
- Bradshaw,R.A. and Burlingame,A.L. (2005) From proteins to proteomics. *IUBMB Life*, **57**, 267–272.
- Cobos,E.S. et al. (2003) A thermodynamic and kinetic analysis of the folding pathway of an SH3 domain entropically stabilised by a redesigned hydrophobic core. *J. Mol. Biol.*, **328**, 221–233.
- Del Sol,A. et al. (2005) Topology of small-world networks of protein–protein complex structures. *Bioinformatics*, **21**, 1311–1315.
- Dijkstra,E.W. (1959) A note on two problems in connection with graphs. *Numer. Math.*, **1**, 269–271.
- Edwards,A.L. (1976) *An Introduction to Linear Regression and Correlation*. W.H. Freeman, San Francisco.
- Fersht,A.R. (1999) *Structure and Mechanism in Protein Science*. W.H. Freeman, New York.
- Galarneau,A. et al. (2002) Beta-lactamase protein fragment complementation assays as *in vivo* and *in vitro* sensors of protein–protein interactions. *Nat. Biotechnol.*, **20**, 619–622.
- Gilbert,M. et al. (2004) Protein expression arrays for proteomics. *Methods Mol. Biol.*, **264**, 15–23.
- Grantcharova,V.P. and Baker,D. (2001) Circularization changes the folding transition state of the SRC SH3 domain. *J. Mol. Biol.*, **306**, 555–563.
- Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.
- Hanley,J.A. and McNeil,B.J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**, 839–843.
- Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
- Hennecke,J. et al. (1999) Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. *J. Mol. Biol.*, **286**, 1197–1215.
- Hubbard,S.J. et al. (1991) Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.*, **220**, 507–530.
- Ihaka,R.G. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
- Iwakura,M. et al. (2000) Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.*, **7**, 580–585.
- Jones,B.E. and Matthews,C.R. (1995) Early intermediates in the folding of dihydrofolate reductase from *Escherichia coli* detected by hydrogen exchange and NMR. *Protein Sci.*, **4**, 167–177.
- Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
- Lee,C. et al. (2001) Role of the connectivity of secondary structure segments in the folding of alpha(1)-antitrypsin. *Biochem. Biophys. Res. Commun.*, **287**, 636–641.
- Lindberg,M. et al. (2002) Complete change of the protein folding transition state upon circular permutation. *Nat. Struct. Biol.*, **9**, 818–822.
- Murzin,A.G. et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Otzen,D.E. and Fersht,A.R. (1998) Folding of circular and permuted chymotrypsin inhibitor 2: retention of the folding nucleus. *Biochemistry*, **37**, 8139–8146.
- Santonico,E. et al. (2005) Methods to reveal domain networks. *Drug Discov. Today*, **10**, 1111–1117.
- Sternberg,M.J. et al. (1995) Towards an intelligent system for the automatic assignment of domains in globular proteins. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 376–383.
- Topell,S. et al. (1999) Circularly permuted variants of the green fluorescent protein. *FEBS Lett.*, **457**, 283–289.
- Vendruscolo,M. et al. (2001) Three key residues form a critical contact network in a protein folding transition state. *Nature*, **409**, 641–645.
- Vendruscolo,M. et al. (2002) Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **65**, 061910.
- Weikl,T.R. and Dill,K.A. (2003) Folding kinetics of two-state proteins: effect of circularization, permutation, and crosslinks. *J. Mol. Biol.*, **332**, 953–963.
- Xenarios,I. et al. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.