Journal of Complex Networks (2019) **7**, 101–113
doi: 10.1093/comnet/cny014
Advance Access Publication on 8 August 2018

Modular decomposition of protein structure using community detection

William P. Grant[†]
Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK
[†]Corresponding author. Email: wpg23@cam.ac.uk

and

Sebastian E. Ahnert
Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK and Sainsbury Laboratory, University of Cambridge, Cambridge CB2 1LR, UK

Edited by: Ernesto Estrada

[Received on 4 March 2018; editorial decision on 29 June 2018; accepted on 4 July 2018]

As the number of solved protein structures increases, the opportunities for meta-analysis of this dataset increase too. Protein structures are known to be formed of domains; structural and functional subunits that are often repeated across sets of proteins. These domains generally form compact, globular regions, and are therefore often easily identifiable by inspection, yet the problem of automatically fragmenting the protein into these compact substructures remains computationally challenging. Existing domain classification methods focus on finding subregions of protein structure that are conserved, rather than finding a decomposition which spans the full protein structure. However, such a decomposition would find ready application in coarse-graining molecular dynamics, analysing the protein's topology, in *de novo* protein design and in fitting electron microscopy maps. Here, we present a tool for performing this modular decomposition using the Infomap community detection algorithm. The protein structure is abstracted into a network in which its amino acids are the nodes, and where the edges are generated using a simple proximity test. Infomap can then be used to identify highly intra-connected regions of the protein. We perform this decomposition systematically across 4000 distinct protein structures, taken from the Protein Data Bank. The decomposition obtained correlates well with existing PFAM sequence classifications, but has the advantage of spanning the full protein, with the potential for novel domains. The coarse-grained network formed by the communities can also be used as a proxy for protein topology at the single-chain level; we demonstrate that grouping these proteins by their coarse-grained network results in a functionally significant classification.

Keywords: community detection; protein structure; biological networks; spatial networks.

1. Introduction

All proteins are formed of chains of covalently bonded amino acids (also known as residues). The pattern of non-covalent bonding between units of the chain is what causes the protein to fold into its compact native structure; specifying the sequence of amino acids in a protein is sufficient to uniquely determine its folded shape [1]. This structure then allows the protein to carry out its designated role within the cell.

Solving a protein's structure is costly in time and effort, yet the number of solved structures is growing rapidly. Over 130 000 protein structures are now publicly available in the Protein Data Bank (PDB) [2], and the size of this dataset is growing exponentially [3]. A widely-researched option for extracting insight from this dataset involves the search for protein domains; functional or structural subunits of a protein structure. Finding domains that are conserved between proteins helps to elucidate the relationship between a protein's structure and its function in the cell, and to classify the proteins into a taxonomy based upon their common structural features. The first efforts to assign protein domains were based upon manual expert curation [4]. In recent years, two alternative databases involving both manual curation and computational assignment have emerged as mainstays; the CATH [5] and SCOPe [6] databases. These databases focus on the domain as a structurally conserved unit, rather than as a compact, globular substructure, and as such the SCOPe and CATH labellings of the protein do not span the complete structure. Another widely used tool is the PFAM database [7], which uses hidden Markov models to discover conserved regions of protein sequence.

One plausible alternative definition of a domain is that of a community on a protein structure network. Protein structure networks have been widely used in which the protein's amino acids are taken as the nodes of the network, with a wide variety of approaches taken to generate the edges, often using proximity of the C_α atoms (the central atom in each amino acid, bonded both to the amino acid's side chain and to the neighbouring amino acids via peptide bonds) [8]. This abstraction has shown promise in analysing individual proteins to identify key residues (amino acids) in allosteric communication [9–12] and protein thermal stability [13]. Tools have been developed to assist with the creation and visualization of the networks [14, 15].

The community structure of protein structure networks has been previously studied for individual proteins [16–18], showing that the community structure often aligns well with intuitive functional domains. Other work [19, 20] has validated network-based clustering over more traditional spatial clustering methods such as k-means clustering [21] and average-linkage clustering [22].

However, previous network-based methods [19, 20] have yet to be scaled to the set of proteins as a whole, possibly due to the computational cost involved. In this work, we provide a comparison of network communities to known domain assignments for a large set of distinct proteins (4000 non-redundant protein chains). We offer an approach using the Infomap community detection method, which uses the compression of a random walker's movement on the network to detect hierarchical community structure [23]. This notion of hierarchical community structure is required in order to account for the known multi-scale structure of proteins. We introduce a modified Jaccard measure to validate the generated community structures, and investigate the coarse-grained networks obtained by condensing each community into a single node, as a proxy for protein architecture.

Non-network-based comprehensive studies of protein structure such as [22] only compare the numbers of domains found, not the assignments of residue positions to domains. Such approaches would therefore also not allow us to generate condensed networks of modules, and ignore or discard information about the hierarchical nature of community structure, for example by choosing a single cut-off point for the clustering dendrogram [22].

2. Methods

The analysis consists of three steps: the generation of the network from the protein structure, the community detection on the network and the storage and analysis of the communities as regions of the protein.

2.1 *Network generation*

There are many plausible approaches to generating a network representation of a protein's structure. The nodes of the network could be either the protein's atoms [11] or residues [8]. For a residue network, the edges are generated if two residues are within a certain distance. This distance measure can be based upon the inter-C_α distance, the inter-C_β distance, or on the number of pairs of atoms within a certain proximity. Previous literature [8, 14] has established a cut-off distance of 8 Å for C_α or C_β networks, and $\sim$ 5 Å for networks based on the number of neighbouring atoms.

Here, a naïve yet flexible approach to network generation is used, which can generate either atomic networks or residue networks as required. Given the atomic positions from a PDB file, we let the atoms be nodes of the network. Undirected edges are then generated between atoms that are closer than a given cut-off distance. The cut-off distance between atoms i and j is defined as $c_{ij} = s\left(r_i + r_j\right)$, where r_i is the covalent radius of atom i, and s is a scaling parameter that can be varied to generate a network with higher or lower connectivity as required. If an atomic network is required, the edges are linearly weighted by proximity of the relevant atoms. If a residue network is required, the network is condensed by letting the amino acids be nodes in the network, with edges weighted according to the number of neighbouring atoms in the original atomic network. In what follows, residue networks with a value of $s = 4$ are used, following [8].

Performing this analysis on a protein with multiple chains often results in a network with distinct connected components, corresponding to each chain. As such, for this analysis the proteins are first split by chain. This helps ensure that any results are fixed at the sub-quaternary level.

Using a network generation tool written in Rust [24], PDB files containing 10 000 atoms can be parsed in this way in under 1 s.

2.2 *Community detection*

In choosing a community detection algorithm, we require a method that does not require the length scale or number of communities to be specified beforehand; we also require a method that is fast enough to allow for all 130 000 proteins in the PDB to be analysed in a reasonable timeframe. We need the method to detect hierarchical community structure, in order to investigate the multi-scale structure of the protein, and a method with a resolution limit that will not impede the discovery of domain-level structure. Infomap [23] satisfies these constraints, along with known accuracy on benchmark graphs. Infomap has the disadvantage that it is prone to overpartitioning networks with geometric constraints, including spatial networks such as those generated in this work [25]. However, empirically we see that the partitions generated correspond well to the domain-level structure of the protein (see overleaf).

2.3 *Storage*

All networks, partitions and results are stored in a MongoDB database [26]. This prevents duplication of effort; for a given parameter set, the database is first queried for the relevant information. If not found, then the relevant calculation is performed and the results stored in the database. In this way a large data set of protein structures with their community structure can be acquired.

2.4 *Performance*

In order to compare the match between the structure found using community detection and that found using other methods, we need a quantitative measure of similarity [27]. Traditional performance metrics such as the Normalized Mutual Information are unsuitable for this task; the predicted structure (for instance the PFAM domain structure [7]) generally occupies only a subset of the protein, whilst the generated community structure tiles the protein completely. Extra structure outside the region spanned by the prediction should not be penalized.

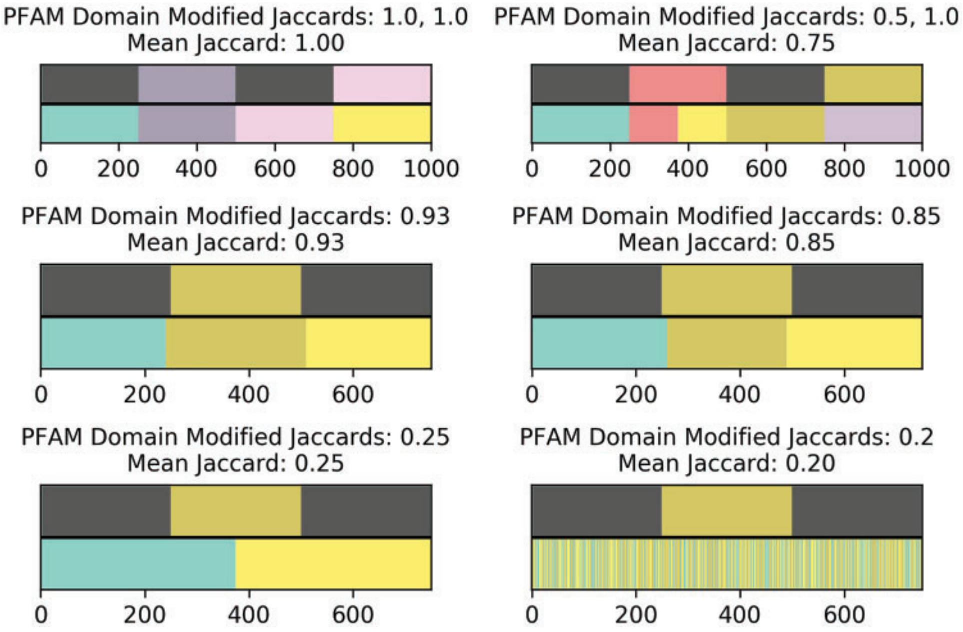


FIG. 1. An illustration of the modified JI on example data. The expected domains (for instance, PFAM domains) are given above, and example community structures below for six possible cases. Each coloured block indicates a domain, with grey indicating unannotated regions. The upper right example shows a perfect score, as each PFAM domain matches a community perfectly. In contrast, the upper left figure shows that one PFAM domain has been split into two communities, giving the matching to that domain a score of 0.5 and an average score for the total match of 0.75. Note that the lower right figure, representing roughly the poorest imaginable case (a randomly shuffled two-community partition), still achieves a modified JI of 0.2.

To this end, we modify the Jaccard index (JI), as follows. The JI is defined as the intersection between two sets, divided by their union, where in this case the sets correspond to regions of the protein sequence. This index is modified as follows:

For each ‘expected’ domain:

- Calculate the JI for all generated communities that overlap with the expected domain, i.e. $\frac{A\cap B}{A\cup B}$, where $A \cap B$ is the size of the overlap and $A \cup B$ the total length of sequence spanned by either the expected domain or the generated community.
- Perform an average of all the calculated JIs, weighted by the proportion of the total expected domain spanned by each community.

This gives a score for each expected domain in the protein, indicating how well it is reflected in the community structure. On test data, this modified JI performs reasonably (see Fig. 1), giving high scores to close matches and low scores to poor matches. Note that like the original JI [27], this score does not take values in the full range [0, 1].

In order to calculate the significance of a given modified JI, we use the z-score. This is defined as:

$$z = \frac{\tilde{J} - \mu}{\sigma}$$

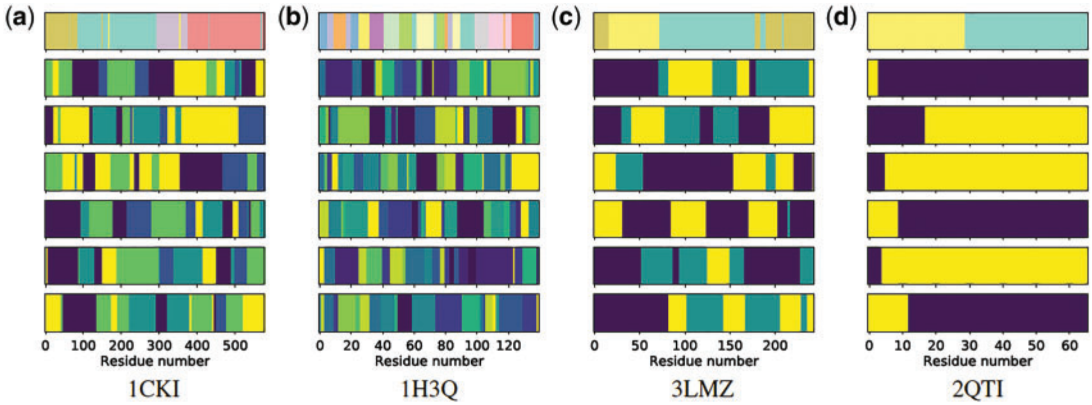


FIG. 2. Randomly generated null models, such that the number of boundaries between communities, and the total number of communities, is preserved. The generated structure to be tested is shown above, with six null models shown below. These null models succeed in capturing the properties of the test structure, ensuring that the comparison between null model and test is fair.

Where $\tilde{J}$ is the modified JI between the expected and generated partitions. μ and σ are the average and standard deviation of the modified JI between the expected partition and a set of null models. μ therefore indicates the modified JI expected by chance. A z-score of two indicates that the modified JI between the generated and expected partitions is two standard deviations higher than the expected value, and therefore, corresponds to a p-value of ~ 0.02 (assuming a normal distribution).

These null models should be randomly generated, sharing some key properties of the generated community structure. In this work, the null models are created by constraining the number of boundaries (changes from one community to another along the sequence), and the total number of communities. Boundaries and community labels are then placed randomly to obey these constraints. Figure 2 shows the community structure to be tested above, with six generated null models below. These models succeed in capturing the rough features of the generated structure, whilst preserving randomness.

3. Results

Empirically, we see that a scaling parameter of approximately 4 gives communities corresponding to compact, globular regions of the protein structure (Fig. 3). We can quantify the extent to which these communities overlap with known protein annotations using the z-score as defined previously. Here, we test the correspondence between the known PFAM domains, and the generated community structure. In general, there is significant agreement, with the majority of proteins having a z-score greater than 2 (Fig. 4).

The communities found are based purely on the protein's structure, whilst the PFAM domains are based purely on sequence. As such, we expect discrepancies when the PFAM sequence domains correspond to more spatially extended, less well-connected regions of the structure. We can measure this by calculating the conductance of the regions of the network responding to the PFAM domains. If the set of nodes of a network V is split into two subsets S and $\bar{S}$, the conductance is defined as:

$$\phi(S) = \frac{\sum_{i\in S, j\in \bar{S}} A_{ij}}{\min(A(S), A(\bar{(S)}))}$$

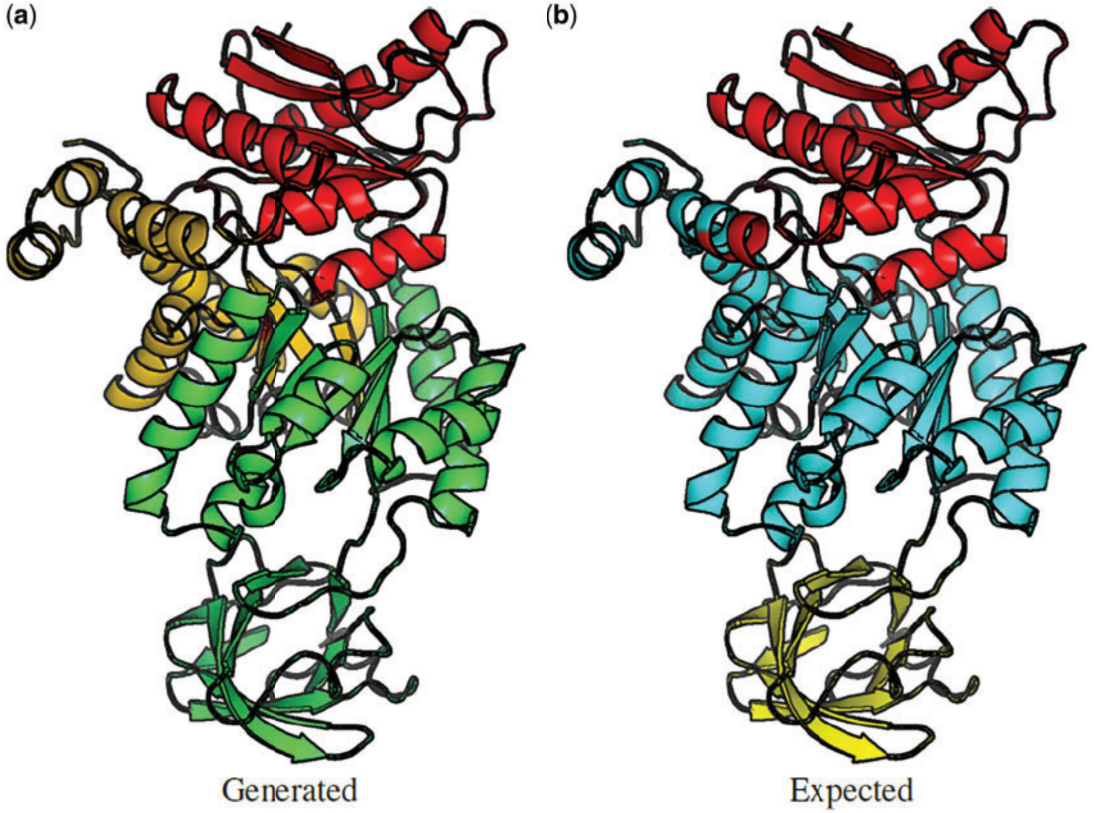


FIG. 3. The generated community structure (a) using a residue network with scaling parameter $s = 4$ compared with the known SCOPe domains (b) for a pyruvate kinase with PDB code 1PKN. One colour signifies one domain/community. Here, we see that one of the communities matches well to the existing SCOPe domain (both shown in red).

Where A_{ij} are the elements of the networks adjacency matrix, and $A(S) = \sum_{i \in S} \sum_{j \in V} A_{ij}$. Hence $\phi(S) \in [0, 1]$, with a lower conductance corresponding to a more isolated region of the network. We expect the modified JI and the conductance to negatively correlate; Figure 5 shows this is indeed the case.

We can compare the communities generated using Infomap to previous network-based attempts to assign domains, which used correlation networks and a modularity-based method [20]. Figure 6 compares these results qualitatively to SCOPe annotations and to the results obtained using our protocol. Figure 7 compares the results quantitatively, using the z-score. A drawback of the correlation-based approach is that a set of homologous proteins is needed; our method has the advantage that it can be performed on single proteins, meaning that the partition spans the full protein structure, and making the approach scalable to larger datasets.

In addition to the communities' potential value as structural domains, the arrangement of the communities may be used as a proxy for topology. The community structure can be converted to a coarse-grained network in which the protein's communities become nodes, linked if the respective communities are neighbours. We can then classify the proteins according to the arrangement of their communities, by grouping proteins with isomorphic coarse-grained networks.

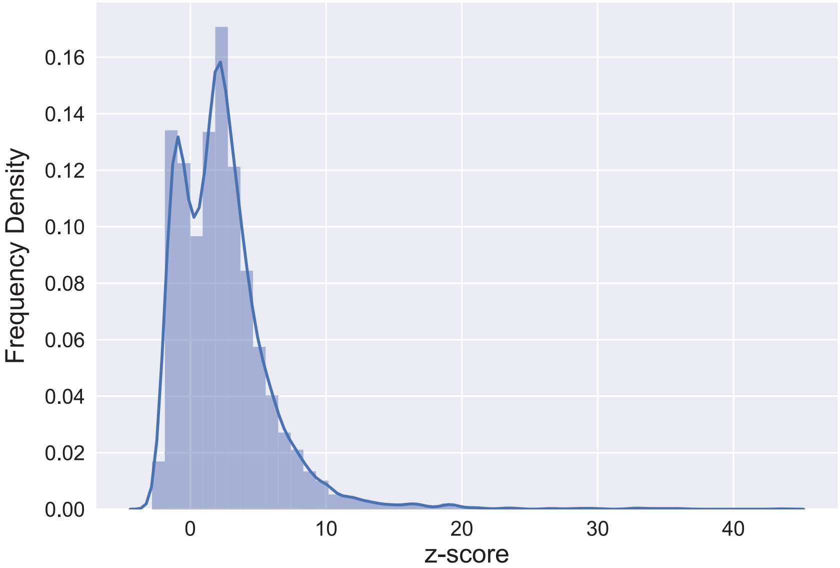


FIG. 4. Histogram showing the z-score for the modified JI between the generated community structure, and the PFAM domains, for ~1000 test proteins, showing that in many cases the agreement is extremely significant.

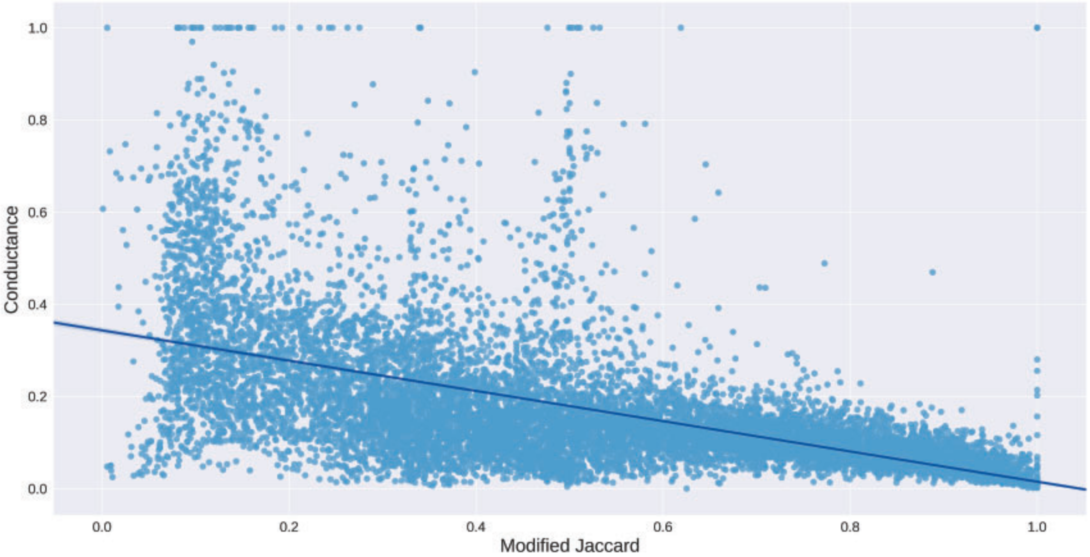


FIG. 5. The conductance of the PFAM domain, when mapped onto the network, against the modified JI (indicating how well it corresponds to the community structure) The conductance is 0 for perfectly-isolated communities, and 1 for communities fully connected to the rest of the network, so we expect a negative correlation between modified JI and conductance; this is seen for the proteins studied here.

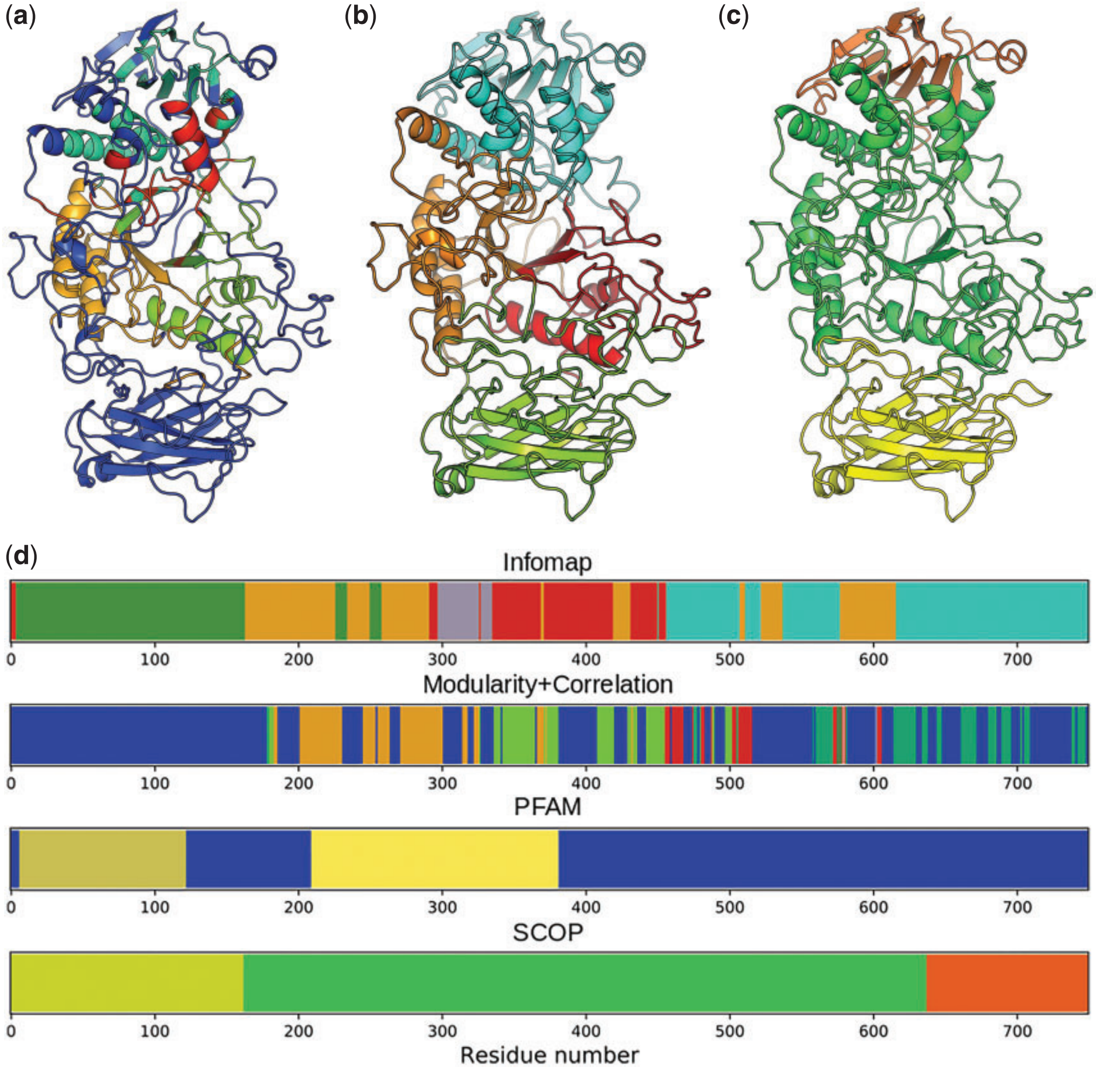


FIG. 6. A comparison of annotations of the protein 1BF2. (a) The decomposition generated in previous work using a modularity-based method, combined with residue correlation analysis [20]. Dark blue regions correspond to unannotated regions of the structure. (b) The decomposition using Infomap presented in this article. (c) The domains listed in the SCOPe structural domain database. (d) The same comparison, along with the PFAM annotations, presented as labellings of the protein sequence. Again, dark blue represents unannotated regions of the sequence.

If the community structure is truly capturing the protein's topology, we expect this grouping to reveal aspects of protein function. We can test this claim using Gene Ontology (GO) term analysis [28]. This effort assigns functional relevance (e.g. lactase activity, oxidoreduction) to genes. The SIFTS project [29] maps these GO terms to records in the PDB, meaning that each protein now has a set of labels encoding information about its function in the cell. We can then test if the grouping results in enriched GO terms,

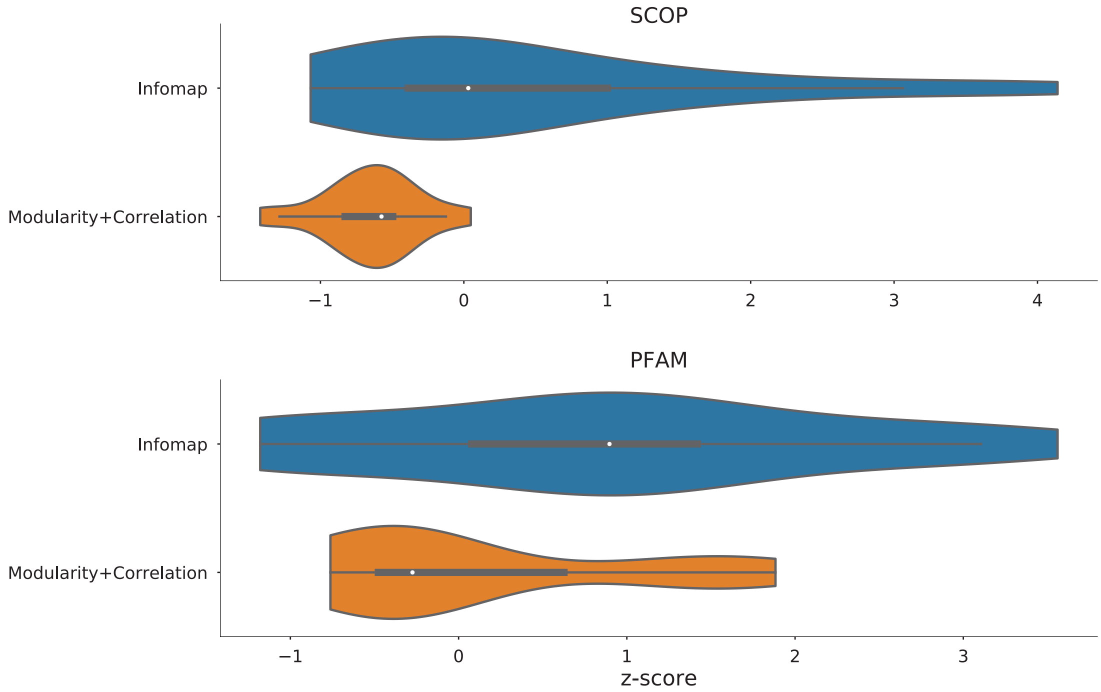


FIG. 7. A comparison of the z-scores for a set of 20 reference proteins, giving the significance of the overlap between SCOP (above) and PFAM (below) for the modularity + correlation method [20] and the proposed Infomap-based method. We see that in both cases the Infomap-based method has a more significant similarity to existing annotations.

i.e. terms appearing more often than expected by chance [30, 31]. For N total proteins, and a subset of that dataset with n proteins, the probability of a GO term being found is given by the cumulative distribution function (CDF) of the hypergeometric function. For a given GO term, let k be the number of times it occurs in the subset, and K be the number of times it occurs across the full dataset. Then the likelihood that the term would be seen k times by chance is:

$$CDF = 1 - \frac{\binom{n}{k+1}\binom{N-n}{K-(k+1)}}{\binom{N}{K}} \, {}_3F_2\left[\begin{matrix}1, & & k+1-K, & & k+1-n \\ & k+2, & & N+k+2-K-n & \end{matrix} ;1\right]$$

Where ${}_3F_2$ is the generalized hypergeometric function. From the CDF, we can acquire p-values for a given grouping and GO term; we consider GO-terms with a p-value of less than 0.01 to be enriched in the subset to a statistically significant extent. As we testing M distinct GO terms, we account for multiple hypothesis testing by applying the Bonferroni correction. If comparisons of M GO terms are being made, the raw p-value is multiplied by M to give a more conservative estimate of the likelihood.

We see that 90% of the proteins studied can be represented by only 10 coarse-grained networks, all of which are associated with GO-term enrichment (Fig. 8 and Table 1).

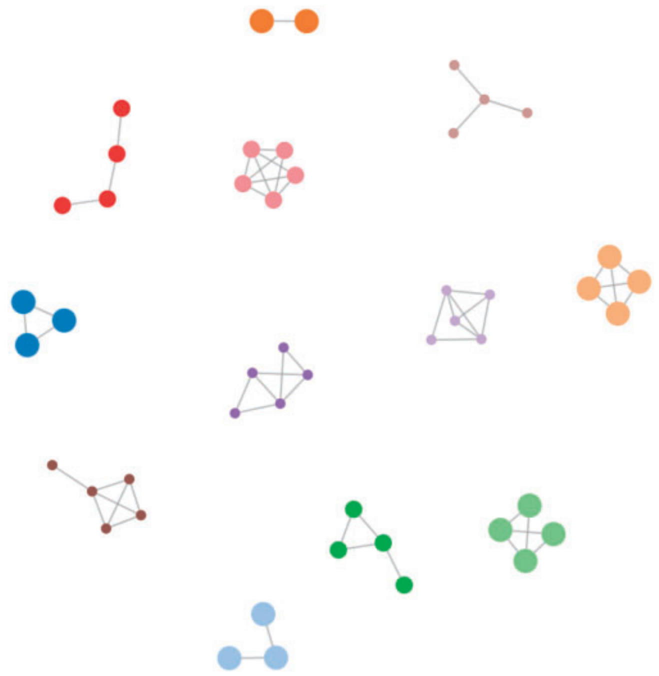


FIG. 8. The coarse-grained networks generated from the community structures of approximately 4300 protein chains, taken randomly from a non-redundant subset of the Protein Data Bank. Only coarse-grained networks common to at least three proteins are shown (accounting for $\sim$ 3900 proteins in total). The node size is proportional to the number of proteins exhibiting that coarse-grained network.

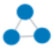
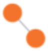
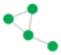
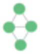
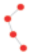
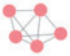
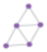
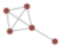
4. Conclusion

There have been many attempts to define the domain, as a compact, repeated unit of protein structure. But choosing these compact, globular substructures in an automated way has traditionally been challenging. We present results showing that a simple weighted network of residue contacts analysed with Infomap can successfully fragment a protein into compact modules. By using a modified JI, we show that in general these modules correlate well with existing PFAM annotations, yet have the advantage that they span the full protein structure. This has potential applications in molecular dynamics and electron microscopy.

We also show that by generating a coarse-grained network, in which the communities of the network are taken as nodes, we can group a large set of proteins in a way that gives significant functional enrichment, as measured by the prevalence of GO terms. This suggests that the community structure can be used as a proxy for the protein topology.

The next step will be to use this approach to search for repeated communities with similar internal topology that have not yet been identified as domains, with the hope of establishing a new framework for domain discovery.

Table 1 *The ten most common protein topologies in the dataset studied, ordered by prevalance*

Coarse-grained network	Number of enriched GO terms ($p<0.01$)	Number of proteins
	331	1725
	130	307
	116	841
	104	445
	34	55
	52	207
	48	26
	23	19
	9	6
	22	8
	7	4

Funding

This work was supported by the Engineering and Physical Sciences Research Council Centre for Doctoral Training in Computational Methods Materials Science [EP/L015552/1] to W.P.G. and the Royal Society and the Gatsby Foundation to S.E.A.

References

1. Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
2. Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., Costanzo, L. D., Duarte, J. M., Dutta, S., Feng, Z., Green, R. K., Goodsell, D. S., Hudson, B., Kalro, T., Lowe, R., Peisach, E., Randle, C., Rose, A. S., Shao, C., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J. D., Woo, J., Yang, H., Young, J. Y., Zardecki, C., Berman, H. M. & Burley, S. K. (2016) The RCSB protein data bank: integrative view of protein, gene and 3d structural information. *Nucleic Acids Res.*, **45**, D271–D281.
3. Berman, H. M., Coimbatore Narayanan, B., Costanzo, L. D., Dutta, S., Ghosh, S., Hudson, B. P., Lawson, C. L., Peisach, E., Prlić, A., Rose, P. W., Chenghua, S., Huanwang, Y., Jasmine, Y. & Christine, Z. (2013) Trendspotting in the protein data bank. *FEBS Lett.*, **587**, 1036–1045.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, **247**, 536–540.

5. Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A. & Sillitoe, I. (2017) CATH: an expanded resource to predict protein function through structure and sequence, *Nucleic Acids Res.*, **45**, D289–D295.

6. Fox, N. K., Brenner, S. E. & Chandonia, J.-M. (2013) SCOPe: structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.

7. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

8. Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G. & Shen, B. (2014) The construction of an amino acid network for understanding protein structure and function. *Amino Acids*, **46**, 1419–1439.

9. Del Sol, A., Araúzo-Bravo, M. J., Amoros, D. & Nussinov, R. (2007) Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome Biol.*, **8**, R92.

10. Di Paola, L. & Giuliani, A. (2015) Protein contact network topology: a natural language for allostery. *Curr. Opini. Struct. Biol.*, **31**, 43–48.

11. Amor, B. R., Schaub, M. T., Yaliraki, S. N. & Barahona, M. (2016) Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. *Nature Commun.*, **7**, article number 12477.

12. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. & Pietrokovski, S. (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, **344**, 1135–1146.

13. Csermely, P., Korcsmáros, T., Kiss, H. J., London, G. & Nussinov, R. (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.*, **138**, 333–408.

14. Chakrabarty, B. & Parekh, N. (2016) NAPS: Network analysis of protein structures. *Nucleic Acids Res.*, **44**, W375–W382.

15. Doncheva, N. T., Klein, K., Domingues, F. S. & Albrecht, M. (2011) Analyzing and visualizing residue networks of protein structures. *Trends Biochem. Sci.*, **36**, 179–182.

16. Delvenne, J. C., Yaliraki, S. N. & Barahona, M. (2010) Stability of graph communities across time scales. *Proc. Natl. Acad. Sci. USA*, **107**, 12755–12760.

17. Delmotte, A., Tate, E. W., Yaliraki, S. N. & Barahona, M. (2011) Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin–myosin light chain interaction. *Phys. Biol.*, **8**, 055010.

18. Zhang, H., Salazar, J. D. & Yaliraki, S. N. (2017) Proteins across scales through graph partitioning: application to the major peanut allergen Ara h 1. *J. Complex Netw.*, cnx052.

19. Tasdighian, S., Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., Palumbo, P., Mei, G., Di Venere, A. & Giuliani, A. (2013) Modules identification in protein structures: the topological and geometrical solutions. *J. Chem. Inf. Model.*, **54**, 159–168.

20. Hleap, J. S., Susko, E. & Blouin, C. (2013) Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. *BMC Struct. Biol.*, **13**, 20.

21. Jain, A. K. (2010) Data clustering: 50 years beyond k-means. *Pattern Recog. Lett.*, **31**, 651–666.

22. Feldman, H. J. (2012) Identifying structural domains of proteins using clustering. *BMC Bioinformatics*, **13**, 286.

23. Rosvall, M. & Bergstrom, C. T. (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS One*, **6**, 1–10.

24. Matsakis, N. D. & Klock, F. S. (2014) II, The rust language. *Ada Lett.*, **34**, 103–104.

25. Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N. & Barahona, M. (2012) Markov dynamics as a zooming lens for multiscale community detection: non clique-like communities and the field-of-view limit. *PloS One*, **7**, e32210.

26. Chodorow, K. (2013) *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*. Sebastopol, C.A.: O'Reilly Media, Inc.

27. Fortunato, S. & Hric, D. (2016) Community detection in networks: a user guide. *Phys. Rep.*, **659**, 1–44.

28. ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

29. VELANKAR, S., DANA, J. M., JACOBSEN, J., VAN GINKEL, G., GANE, P. J., LUO, J., OLDFIELD, T. J. ODONOVAN, C., MARTIN, M.-J. & KLEYWEGT, G. J. (2013) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.

30. HUANG, D. W., SHERMAN, B. T. & LEMPICKI, R. A. (2008) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

31. RHEE, S. Y., WOOD, V., DOLINSKI, K. & DRAGHICI, S. (2008) Use and misuse of the gene ontology annotations. *Nature Rev. Genet.*, **9**, 509–515.