

Proteins across scales through graph partitioning: application to the major peanut allergen Ara h 1

HENG ZHANG

Department of Chemistry, Imperial College London, London SW7 2AZ, UK

JOSÉ DOMINGO SALAZAR

Imperial College London, London SW7 2AZ, UK

AND

SOPHIA N. YALIRAKI[†]

Department of Chemistry, Imperial College London, London SW7 2AZ, UK

[†]Corresponding author. Email: s.yaliraki@imperial.ac.uk

Edited by: Hocine Cherifi

[Received on 18 March 2017; editorial decision on 11 October 2017; accepted on 13 October 2017]

The analysis of community structure in complex networks has been given much attention recently, as it is hoped that the communities at various scales can affect or explain the global behaviour of the system. A plethora of community detection algorithms have been proposed, insightful yet often restricted by certain inherent resolutions. Proteins are multi-scale biomolecular machines with coupled structural organization across scales, which is linked to their function. To reveal this organization, we applied a recently developed multi-resolution method, Markov Stability, which is based on atomistic graph partitioning, along with theoretical mutagenesis that further allows for *hot spot* identification using Gaussian process regression. The methodology finds partitions of a graph without imposing a particular scale *a priori* and analyses the network in a computationally efficient way. Here, we show an application on peanut allergenicity, which despite extensive experimental studies that focus on epitopes, groups of atoms associated with allergenic reactions, remains poorly understood. We compare our results against available experiment data, and we further predict distal regulatory sites that may significantly alter protein dynamics.

Keywords: peanut allergenicity; major peanut allergen Ara h 1; community detection; Markov Stability; all-scale method; protein dynamics; mutational analysis.

1. Introduction

Proteins are multi-scale biomolecular machines with coupled structural organizations across time and length scales. The global dynamics of a protein structure arises as a result of contributions from the smallest scale of atoms, through chemical groups, amino acids residues, secondary structures, to the largest scale of domain motions. Each level of this structural organization links to some functional behaviour and the related scales span across several orders of magnitude [1, 2]. This coupling poses challenges to computational methods [3–6] not only due to their excessive computation cost but also due to sampling issues or appropriate level of description. Coarse-grained methods have been proposed as a means to reach the functionally relevant scales [7–12] but with a cost of loss of generality and of smearing the detailed physicochemical atomic interactions.

Understanding protein dynamics across scales is especially important in a problem such as peanut allergy, an immunoglobulin E (IgE)-mediated hypersensitivity that is triggered by peanut allergens, leading to various physical symptoms including eczema, asthma, diarrhoea and cardiac arrest [13]. The allergy is more common in developed countries than in the developing world. For example, approximately over 1% of the United States population exhibit peanut allergies [14], whereas in England, there are an estimated five thousand people who are diagnosed with this allergy for the first time every year [15].

Even though at least 11 protein allergens have been described with crystallographic structures, the mechanism for eliciting a reaction to showing disease symptoms is poorly understood. There is currently no methodology that can *a priori* robustly predict the allergenicity of a protein. At the same time, proteins with similar structures do not exhibit allergenic behaviour. Ara h 1, a major peanut allergen, is a homotrimeric bicupin protein, held together by hydrophobic interactions at monomer interfaces [16]. Each subunit is readily digestible, as there are no disulphide bonds within [16]. It has been proposed that since peanut allergens should be intact before they reach the intestinal mucosa, there should be a relation between allergen stability and their allergenicity [17, 18]. Follow-up studies focused on the protein monomer before and after digestion by proteases, a special group of proteins that cut the targets into smaller fragments. At least 23 fragments, termed epitopes, were identified [19] as able to bind IgE to trigger allergenicity *in vitro*. Among these fragments, epitopes 8 and 14 are identified as the immunodominant epitopes, as they can bind to most of the IgE samples. Site-directed mutagenesis was performed on these epitopes, and it was found that point mutations at certain residues have strong impact on their binding affinity [20]; these residues are consequently called *critical residues*. Figure 1A summarizes these experimental findings by showing the crystal structure of the core of the protein Ara h 1, with the major immunogenic epitopes and *critical residues*.

Ara h 1 is part of the cupin superfamily of proteins, one of the most functionally diverse families with unusually thermostable proteins and, as such, shares structural similarities with a variety of proteins. However, proteins that are structurally similar to allergens do not exhibit allergenic potential [22, 23] and distinguishing allergens from negative controls remains difficult. For example, oxalate decarboxylase (OxuC), a manganese-dependent enzyme that decomposes oxalate into formate and carbon dioxide, shares 40% sequence similarity and 68.6% structural similarity with Ara h 1. Another negative control, MnCA (Mn^{2+} -cupin A), is 21% sequentially and over 68.4% structurally similar to Ara h 1 (see Fig. 1 and Table 1). Neither has been shown to trigger allergenicity so far even though they are structurally clustered together within their superfamily [24]. Understanding peanut allergens at the molecular level and differentiating the dynamics among similar proteins is therefore important for shedding light to the problem of peanut allergenicity.

A variety of experimental and computational methods have been developed to extract information from allergens and infer the allergenic potential of different proteins. For example, microfluidics diagnostics techniques [25–27] were developed as less time-consuming and cost-effective compared with the double-blind, placebo-controlled food challenge, the gold standard for diagnosing peanut allergy [28–30]. Bioinformatics tools are increasingly utilized to evaluate the similarity between known allergens and a new protein within the food safety assessment process [31–36]. Protein similarity networks are used to elucidate the sequence/structure/function relationships within the cupin superfamily, but differentiating similar proteins based on their allergy-triggering capability is still difficult [24].

In this article, we address the question of peanut allergenicity from a multi-scale dynamic perspective, starting with atomistic descriptions. We use Markov Stability, a recently developed all-scale graph-theoretical method, to determine the intrinsic dynamics of the core of the major peanut allergen Ara h 1 [37]. This method has been successfully applied in studies including protein function domain detection [38, 39] and allostery [40] among others. Compared to previous approaches [41–44], it directly calculates

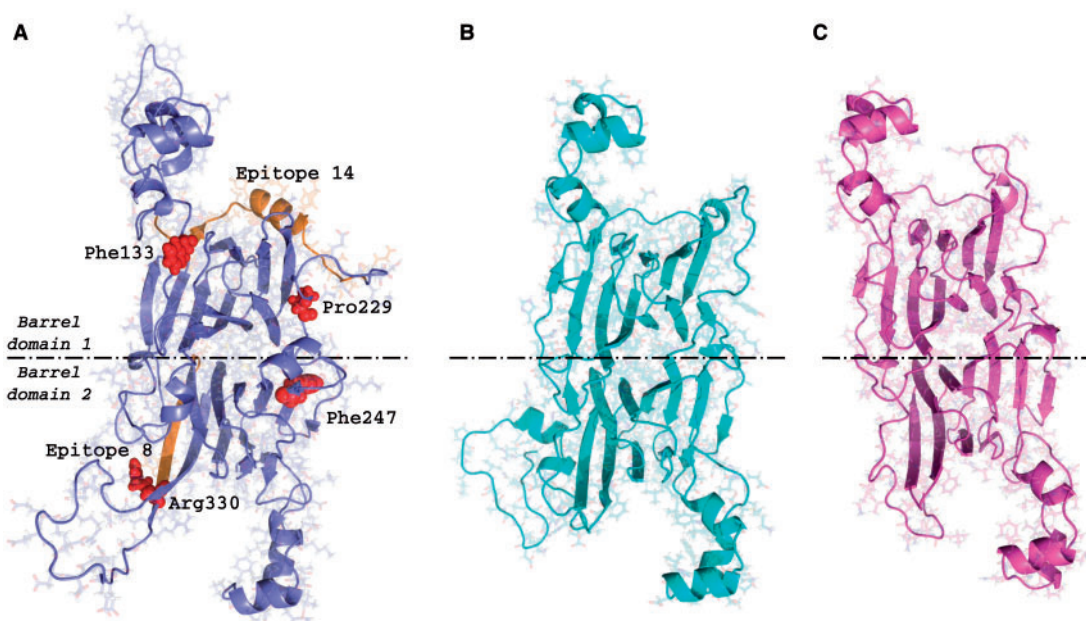


FIG. 1. (A) The crystal structure of the core domain of the major peanut allergen Ara h 1 (PDB ID: 3S7E). The two barrel domains are indicated by the dashed line. The immunodominant epitopes 8 and 14 are highlighted in orange, whereas critical residues Phe133, Pro229, Phe247 and Arg330 are coloured in red. (B) The crystal structure of oxalate decarboxylase (OxdC, PDB ID: 1UW8). (C) The crystal structure of MncA (PDB code: 2VQA). Although the three proteins come from the same family and share structural similarity, including the bifold barrel, only (A) is an allergen.

TABLE 1 *Sequence and structural similarity between Ara h 1 and negative controls, OxdC and MnCA, calculated using the DaliLite v.3 server [21] for pairwise comparisons*

Allergen	Control	Sequence similarity (%)	Structural similarity (%)
Ara h 1	OxdC	40	68.6
	MnCA	21	68.4

different levels of organization in a protein structure across multiple time scales through partitioning while keeping atomistic details in a computationally efficient manner. First, it constructs an atomic-level graph for each protein from its structural data, where the graph nodes are the atoms, and the edge weights are based on interatomic interactions deduced from atomic positions and geometries [38, 40]. It then uses a diffusion on the graph to find communities at each scale by evaluating the quality function defined by Markov Stability [37, 45]. This avoids the resolution problem inherent in other community detection methods [46, 47] that are embedded within a specific scale [39]. Finding communities without imposing a scale *a priori* can help the understanding of a system's global behaviour formed by components existing at various resolutions. The method is also capable of identifying dynamical differences between similar structures [48, 49], so it is well suited for unravelling the dynamical differences of the major peanut allergen Ara h 1 and its negative controls. We first explore the functional components of Ara h 1 through

its structural hierarchy across scales. We show that there is a correspondence between partitions at intermediate time scales and experiment-proposed epitopes (groups of residues thought to be responsible for triggering an allergic reaction) and also between partition boundaries and *critical residues*. We highlight the different partitioning processes of the allergen and two negative controls, OxA and MnCA, as a means to shed light in their differing functions. In parallel, we use a theoretical mutagenesis method to identify *hot spots*, residues that impact the global dynamics of the peanut allergen.

2. All-scale protein dynamics through Markov Stability

2.1 Graph construction from protein crystal structures

We first derive an undirected, weighted atomic network based on the protein structure to be analysed. We briefly summarize the process here, but for details, please see references [40] and [38]. The protein structures are obtained from the RCSB Protein Data Bank (see Fig. 1 and Table 1). Missing residues are completed by the software *MODELLER* [50]. Hydrogen atoms are all stripped and then readed using *REDUCE* [51]. Each node of the network corresponds to an atom and edges correspond to pairwise interactions such as covalent bonds, hydrogen bonds, salt bridges, hydrophobic tethers, electrostatic interactions and π -stacking interactions. Edges are identified using *FIRST* [52] with a cut-off of 0.01 kcal/mol for hydrogen bonds and 8 Å for hydrophobic tethers. They are subsequently weighted by the strength of interactions as follows: the Mayo potential [53] for hydrogen bonds, the potential of mean force from reference [54] for hydrophobic tethers. In this study, we stripped off all ions and water molecules.

2.2 Markov Stability

To reveal the organization of the structure at different scales, we use Markov Stability [45, 55] to analyse the generated protein graph. At each Markov time, we find an optimized partition where a random walk is likely to remain trapped. As we increase this timescale, significant partitions are obtained in increasingly coarser communities. In the case of the protein graph [38, 39, 56], this process allows us to scan across resolutions and find clusters corresponding to chemical groups at very short Markov times, through biochemical units, amino acids and secondary structures, at intermediate Markov times, to partitions corresponding to functional domains or subunits at longer Markov times. These groupings correspond to groups of atoms moving coherently. The duration of the partition, together with its robustness to perturbation, allows us to map out dynamical properties of the protein.

Let us define an undirected, weighted network $G(V, E)$, denoted by the adjacency matrix A of rank n . Its vertex degrees are $d_i = \sum_{ij} A_{ij}$, and the degree matrix $D = \text{diag}(\mathbf{d})$. On this network, we consider a continuous-time Markov process which is governed by the *combinatorial Laplacian*, $L = D - A$, as most appropriate for protein dynamics:

$$\frac{d\mathbf{p}}{dt} D^{-1} \langle d \rangle = -\mathbf{p} D^{-1} [D - A] \Rightarrow \frac{d\mathbf{p}}{dt} = -\frac{1}{\langle d \rangle} \mathbf{p} L, \quad (2.1)$$

where $\langle d \rangle = (\mathbf{1}^T D \mathbf{1})/N$ is the average degree and $\mathbf{p} = \mathbf{p} D^{-1}$. Now the stationary distribution of (2.1) is the uniform distribution over all the nodes: $\pi_c = \mathbf{1}^T/N$.

Markov Stability is then defined as

$$R(t, P) = \text{Tr}[\mathbf{H}^T (\Pi_c e^{-tL/\langle d \rangle} - \pi_c^T \pi_c) \mathbf{H}], \quad (2.2)$$

where c is the number of communities in the partition; $\mathbf{H}$ is a $N \times c$ indicator matrix of P with entries H_{ij} equal to one if node i belongs to community j and zero otherwise; π_c denotes the stationary distribution defined above and Π_c are the diagonal elements of π_c . The time t here is the Markov time or a dimensionless resolution parameter.

2.3 Optimizing Markov Stability using the Louvain algorithm

As is the case in most clustering-related problems, the global optimization of Markov Stability is NP-hard [37]. A variety of heuristic strategies can be used to obtain good partitions that can then be ranked by Markov Stability to provide us with near-optimal partitions at different time scales.

Here, we use the Louvain algorithm, a greedy agglomerative method [57]. The algorithm has been observed to require little computational effort and to find partitions close to the optimal solution [57]. Initially, each node is assigned to its own community, then each node is transferred in turn into the neighbouring community, where the increase of Markov Stability is the largest, as long as it increases the Markov Stability value of the overall partition. This step is repeated until no further transfer can increase the Markov Stability. At this point, a new graph of communities is generated. The algorithm repeats these two steps until a coarse-grained graph is obtained, where no further grouping can improve Markov Stability.

The Louvain algorithm is deterministic, but the final solution found depends on the order in which the different nodes are scanned for the grouping step. This initial ordering, which we refer to as the Louvain initial condition, can be chosen at random every time the calculation is executed. Indeed, we will use the variability of the observed solution induced by our random choice of the Louvain initial condition to estimate the robustness of a partition, a measure of its relevance.

2.4 Defining community robustness by variation of information

A distinctive property of a significant community structure should be its robustness to small perturbations. Upon introducing slight variations in the graph itself, the new partition found should be highly similar to the one obtained originally [37, 56, 58, 59]. One way to quantify the extent to which the partitions are affected by the perturbation [59, 60] is using a measure of distance between the solutions found before and after the perturbation. An information-theoretic distance between two partitions can be measured by the variation of information (VI) [61–63], a metric based on the total information that is not shared by two partitions. The random initial conditions of the Louvain optimization algorithm provide us with an ideal perturbation with respect to which we can measure the robustness of the partitions. By optimizing Markov Stability for an ensemble of such initial conditions for each Markov time, we can calculate the VI between all pairs of solutions obtained and compute the average as a measure of the relevance of the solutions obtained at a particular scale. Other perturbations affecting edge weights or the quality function, for example, have been considered in the past and shown to yield similar results [37, 49].

2.5 Identifying hot spots by mutational analysis

Another question of interest is the identification of residues or nodes that significantly impact the protein dynamics when mutated, referred to as *hot spots*. To mimic *in silico* the process of the experimental procedure, alanine scanning mutagenesis, which replaces each residue at a time with an alanine, we mutate each residue to an alanine in turn by removing all edges of weak interactions with respect to its side chain (see Fig. 2). The mutated graph is partitioned using the linearized version of Markov Stability. We then identify the mutations that affect significantly the robustness of partitions using Gaussian

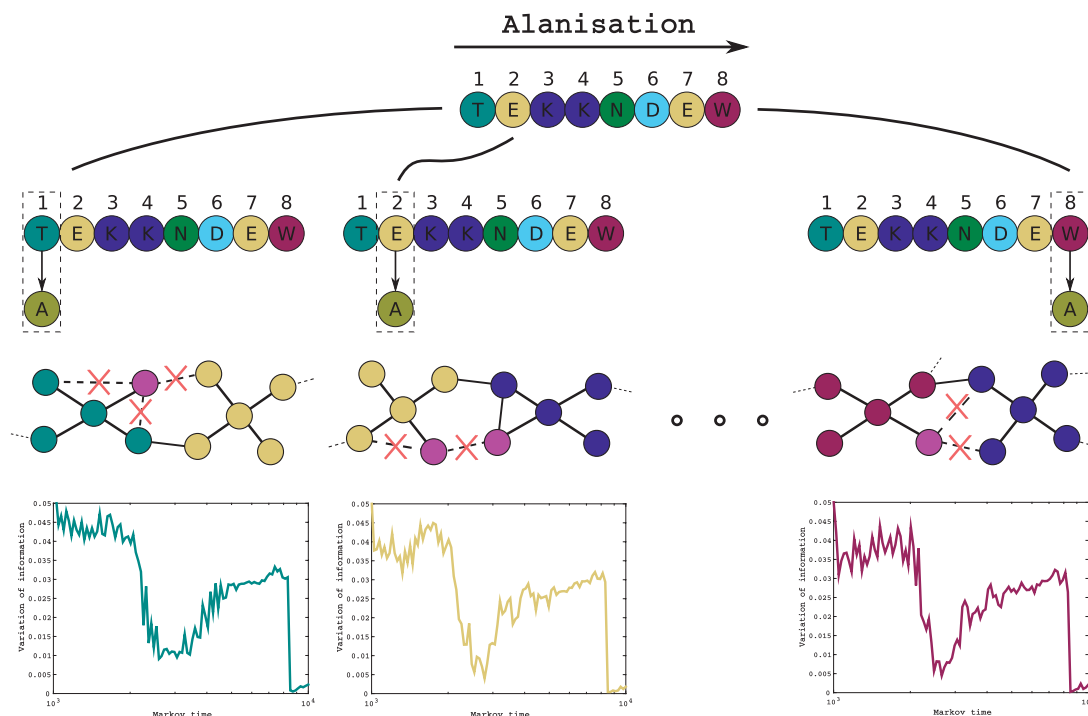


FIG. 2. The mutational analysis, or edge removal, mimics the experimental alanine scanning mutagenesis procedure. The new graph is then analysed with Markov Stability, and the *hot spots* are ranked by applying Gaussian process regression (GPR) [64] to the VI, as explained in the text.

process regression (GPR), a non-parametric fitting method [40, 64]. We use all the VI vectors from the mutation of each residue along with the vector of the wild type to produce a statistically reliable VI vector with confidence bounds. Any mutation with VI values falling outside three standard deviations for more than one-fourth of the time period in the Markov time window is considered to have a significant impact on the graph robustness. The calculation is realized using the gpml MATLAB toolbox (<http://www.gaussianprocess.org/gpml/code/>).

3. Structural organization of the Ara h 1 network across scales

The Markov Stability results of the Ara h 1 monomer are shown in Fig. 3B. The zooming at different resolutions starts by finding chemical groups at high resolution, then onto amino acids and secondary structures, followed by segments and functional domains, and finally merging all parts of the structure. From Markov time 10^5 onwards, Ara h 1 exhibits well-defined communities mostly by long plateaux. At longer time scales, where typically proteins are functional, we observe the long-lived and robust communities in the two-barrel separation. This is in line with how a protein with such an architecture should behave and agrees with our work on other proteins. Interestingly, the intermediate time scales indicate less robustness. For example, the VI of the four-way partition is unusually variable. Additionally, the N-terminus segment between F14 and R17 is partitioned into three consecutive communities. Indeed, it is difficult to find a reasonable partitioning in this region. This lack of partitioning shows that the

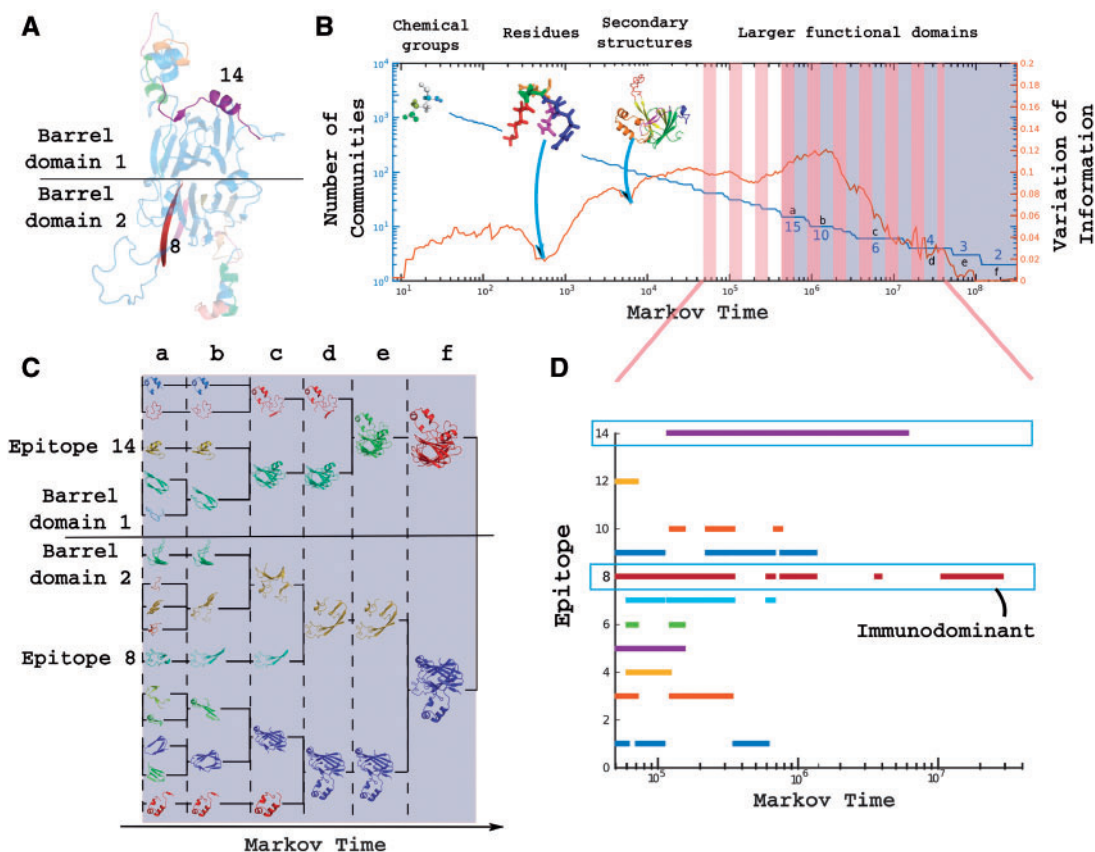


FIG. 3. Structural anatomy of the allergen Ara h 1 at all scales. (A) The atomistic structure of the core of the monomer of Ara h 1 (PDB ID: 3S7E), with the main epitopes identified by Burks *et al.* [19] and Shin *et al.* [20]. (B) Markov Stability analysis of the core of the Ara h 1 monomer. As the Markov time increases, we recover first the meaningful biochemical levels of organization (chemical groups, residues and secondary structures), followed by large segments partly corresponding to reported epitopes and finally partitions of the two large barrels. The detailed community organization of the protein during the intermediate to slow time scales is presented in (C). Eventually the protein partitions in the two-barrel domains. The varying VI is unusual with merged flanking regions. (D) Correspondence between communities across intermediate timescales and epitopes. The most persistent partition is linked with epitope 8, the immunodominant segment.

allergen protein structure is susceptible to local disturbances and may also imply its adaptability to multiple conformations for IgE binding or further aggregation.

Certain communities across those intermediate time scales correlate with some of the proposed epitopes. At each time step, each community was sequentially compared with each of the experimental epitopes by overlapping their atoms. As all Ara h 1 epitopes are linear, communities with breakages were not considered. When one community overlapped with a certain epitope with over 80% of their atoms, a correspondence was established. As the 14th epitope is over twice the size of others, its failing threshold was set to 30%. Figure 3D shows the mapping of epitopes onto the communities in Ara h 1. No community can be mapped with epitope at bigger scales, because smaller communities will merge into large functional domains, indicating global dynamics. Epitopes 8 and 14 last much longer than the

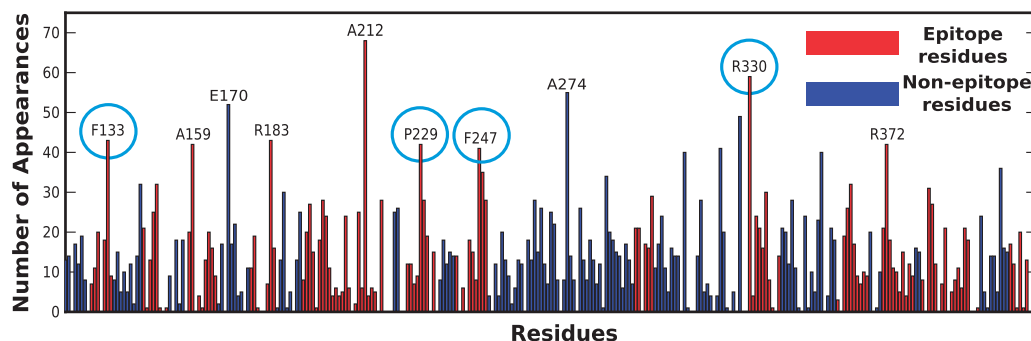


FIG. 4. The highest occurring residues that appear in boundaries of communities at the intermediate time scales are plotted and coloured red if they belong to an epitope and blue otherwise. Residues previously found to be critical are circled. We identify these residues as well but also find additional ones that may play an important role in the protein dynamics.

others, whereas epitopes 2, 11 and 13 did not manifest themselves as single communities. Epitope 8 is an immunodominant epitope according to Burks *et al.* [19], so the persistence of a community is to some extent related to its allergenicity. Note that epitopes appear and reappear due to either merging of communities, for example, epitopes 1 and 3 or breaking of communities of an epitope, for example, epitope 9. As a comparison, there are no linear epitopes identified in OxdC or MnCA.

4. Community boundaries correspond to critical residues

We also mapped community boundary residues at intermediate time scales to epitopes, shown in Fig. 4. We identified four of the epitope residues that have been mutated and known to correspond to *critical residues* that is known to be crucial to IgE binding. However, we also identify further residues that may indicate new pockets for binding events or communication. In particular, residue Ala212 has never been tested experimentally. These residues may be important for functional motions rather than purely thermodynamic or surface exposure reasons.

5. Different partitioning process of the peanut allergen and the controls

Figure 5 shows the Markov Stability analysis results of Ara h 1 and the two controls. As discussed in the Introduction section, the three proteins are structurally similar and share the two-barrelled configuration (see Fig. 1). The two-domain motions are the same when either protein opens and closes around the virtual dyad axis and in fact appear as the final two community partition at the end of the calculation (partition f). However, the evolution of communities is distinct between the allergen and the other structures, reflecting the differing functions they need to perform.

Ara h 1 and OxdC have significantly different number of partitionings even at shorter time scales, indicating different local movements. As time progresses, OxdC formed its C-terminus barrel community first, followed by the other barrel. Then, the C-terminus barrel was split into two, with a varying inner boundary for some time period. When the C-terminus barrel was complete again, it started to merge outside residues, before eventually the two barrels emerged as the two partitions. A similar process was observed in the MncA case. In contrast, the allergen followed a continuous merging of communities until the final two barrels formed (see Fig. 3C).

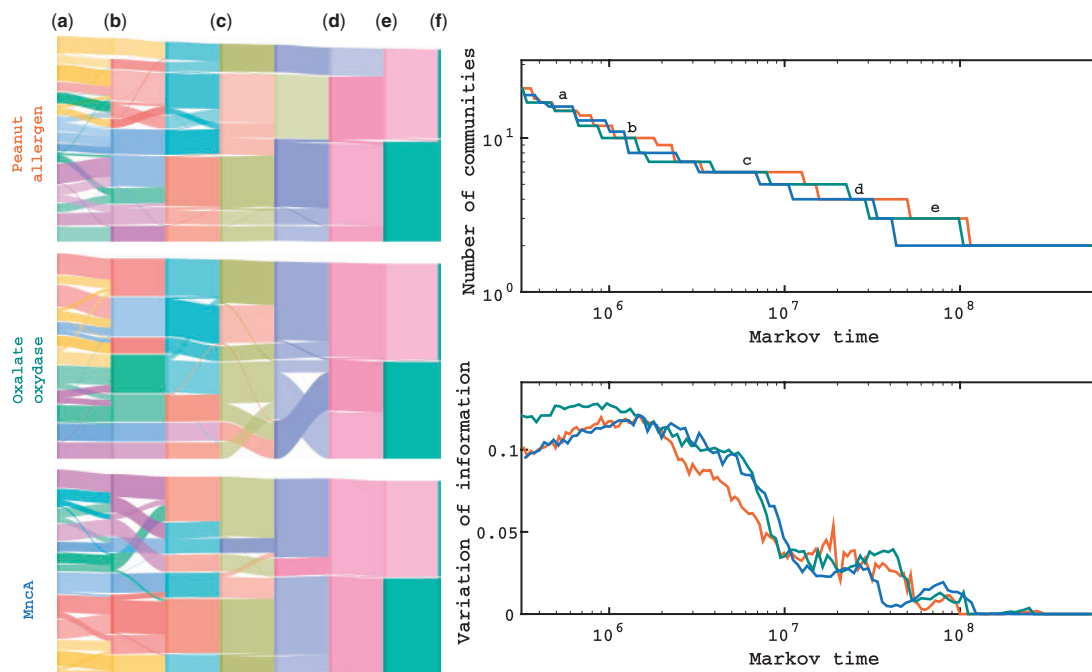


FIG. 5. Markov Stability calculations of the peanut allergen Ara h1 (orange) and the two controls, OxdC (green) and MncA (blue). The number of communities and VI is shown across Markov times. The emergence, duration and robustness of communities differ among the proteins. However, the most striking difference can be observed in the figure on the left, where the community evolution is explicitly shown. In particular, the merging–splitting–merging process of communities is shared by OxdC and MncA, whereas the peanut allergen constantly merges more residues into larger communities, perhaps reflecting the differing functions of the proteins.

In summary, two distinct community evolution processes appear: the merging–splitting–merging process is shared by OxdC and MncA, whereas the peanut allergen constantly merges more residues into bigger communities. These different processes reflect their distinct functions: the allergen, in a consistently co-operative trend, binds IgE at different timescales, while OxdC maintains itself and reorganizes its functional domains to catalyze and cleave carbon–carbon bonds. Most of the critical residues of Ara h1 are distributed on the outside, so the binding process, generally happening on the flanking helices and loops, will not affect the barrel on the inside. In contrast, for OxdC, in addition to its manganese-binding sites positioned in the middle region of each barrel, Just *et al.* [65] argued that Glu162 is a new candidate for the crucial proton donor through substantial conformational change. Consequently, protein segments need to readjust, reflected by the splitting and reforming barrels, which may help explain the different community merging process.

6. Computational site-directed mutagenesis

Figure 6 shows the VI between every mutant of the peanut allergen Ara h1 with the wild type according to the procedure described in the aforementioned Methodology section. Two residues were picked up as having a significant effect by our procedure, namely Glu222 and His211. Glu222 is located by epitope 5, whereas His211 is within epitope 14, beside the partitioning boundary residue Ala212 at medium

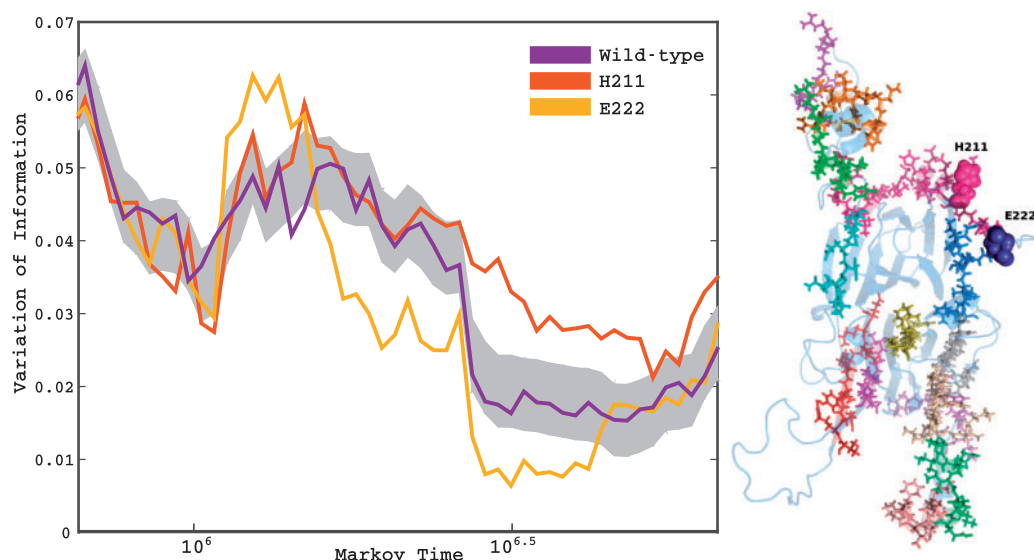


FIG. 6. The VI of the allergen Ara h 1 (purple) and its mutants. Most mutants do not have a significant effect over all timescales and fall within the 95% confidence region by GPR (shown in grey). We consider a residue as a hotspot if its deviation exceeds three standard deviations for over one-fourth of the time spectrum studied. Only mutants H211A (orange) and E222 (yellow) fulfil the criteria.

timescales, and not far from Glu222. As these two residues are directly related to epitopes, it is both their binding affinity and their conformational dynamics that seem to be altered by the mutation.

7. Conclusions and perspectives

Although Ara h 1 is well known for triggering IgE-mediated allergy to release various chemicals, the mechanism of this process is still not fully understood. At least 23 epitopes have been proposed as possible binding places, and the role of epitopes is being re-examined [19].

Here, we studied the major peanut allergen Ara h 1 as well as its structurally similar negative controls through atomistic-based graph partitioning based on Markov Stability as a viable approach for studying protein dynamics across timescales. By partitioning the graphs generated from protein structures, we were able to find meaningful communities at different resolutions related to timescales and functional activities. We identified an intermediate time scale where non-robust communities are related to epitopes, known regions important for allergenic response. We observed distinct coupling routes between levels of dynamics from atomic movements up to functional domains, which may influence the differing functions of IgE-binding activities. Finally, two distal residues had strong impact on the global dynamics when they were mutated by computational alanisation. The extent of the mutational effects and the pathways that may link epitopes are subject of future work.

Acknowledgement

The authors thank Chris Baker, Nathan Kidley, Scott McClain (Syngenta) and Mauricio Barahona (Imperial) for discussions.

Funding

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no 607466 and from the *Engineering and Physical Sciences Research Council (EPSRC)* through award EP/N014529/1 funding the EPSRC Centre for Mathematics of Precision Healthcare.

REFERENCES

1. HENZLER-WILDMAN, K. & KERN, D. (2007) Dynamic personalities of proteins. *Nature*, **450**, 964–972.
2. KUMAR, A. & PUROHIT, R. (2014) Use of long term molecular dynamics simulation in predicting cancer associated SNPs. *PLOS Comput. Biol.*, **10**, 1–14.
3. APOSTOLOVIC, D., STANIC-VUCINIC, D., DE JONGH, H. H. J., DE JONG, G. A. H., MIHAJLOVIC, J., RADOSAVLJEVIC, J., RADIBRATOVIC, M., NORDLEE, J. A., BAUMERT, J. L., MILCIC, M., TAYLOR, S. L., GARRIDO CLUA, N., CIRKOVIC VELICKOVIC, T. & KOPPELMAN, S. J. (2016) Conformational stability of digestion-resistant peptides of peanut conglutins reveals the molecular basis of their allergenicity. *Sci. Rep.*, **6**, 29249–29260.
4. KARPLUS, M. & KURIYAN, J. (2005) Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA*, **102**, 6679–6685.
5. KARPLUS, M. & MCCAMMON, J. A. (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.*, **9**, 646–652.
6. KLEPEIS, J. L., LINDORFF-LARSEN, K., DROR, R. O. & SHAW, D. E. (2009) Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.*, **19**, 120–127.
7. AYTON, G. S., NOID, W. G. & VOTH, G. A. (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.*, **17**, 192–198, Theory and simulation/macromolecular assemblages.
8. BOND, P. J., HOLYOAKE, J., IVETAC, A., KHALID, S. & SANSOM, M. S. (2007) Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.*, **157**, 593–605.
9. DERREUMAUX, P. & MOUSSEAU, N. (2007) Coarse-grained protein molecular dynamics simulations. *J. Chem. Phys.*, **126**, 608–613.
10. INGÓLFSSON, H. I., LOPEZ, C. A., UUSITALO, J. J., DE JONG, D. H., GOPAL, S. M., PERIOLE, X. & MARRINK, S. J. (2014) The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **4**, 225–248.
11. KMIECIK, S., GRONT, D., KOLINSKI, M., WIETESKA, L., DAWID, A. E. & KOLINSKI, A. (2016) Coarse-grained protein models and their applications. *Chem. Rev.*, **116**, 7898–7936.
12. PRONK, S., PÁLL, S., SCHULZ, R., LARSSON, P., BJELKMAR, P., APOSTOLOV, R., SHIRTS, M. R., SMITH, J. C., KASSON, P. M., VAN DER SPOEL, D. HESS, B. & LINDAHL, E. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **7**, 845–854.
13. HONG, X., HAO, K., LADD-ACOSTA, C., HANSEN, K. D., TSAI, H.-J., LIU, X., XU, X., THORNTON, T. A., CARUSO, D., KEET, C. A., SUN, Y., WANG, G., LUO, W., KUMAR, R., FULEIHAN, R., SINGH, A. M., KIM, J. S., STORY, R. E., GUPTA, R. S., GAO, P., CHEN, Z., WALKER, S. O., BARTELL, T. R., BEATY, T. H., FALLIN, M. D., SCHLEIMER, R., HOLT, P. G., NADEAU, K. C., WOOD, R. A., PONGRACIC, J. A., WEEKS, D. E. & WANG, X. (2015) Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. *Nat. Commun.*, **6**, 6304.
14. CABANOS, C., URABE, H., TANDANG-SILVAS, M. R., UTSUMI, S., MIKAMI, B. & MARUYAMA, N. (2011) Crystal structure of the major peanut allergen Ara h 1. *Mol. Immunol.*, **49**, 115–123.
15. KOTZ, D., SIMPSON, C. R. & SHEIKH, A. (2011) Incidence, prevalence, and trends of general practitioner—recorded diagnosis of peanut allergy in England, 2001 to 2005. *J. Allergy Clin. Immunol.*, **127**, 623–630.
16. CHRUSZCZ, M., MALEKI, S. J., MAJOREK, K. A., DEMAS, M., BUBLIN, M., SOLBERG, R., HURLBURT, B. K., RUAN, S., MATTISOHN, C. P., BREITENEDER, H. & MINOR, W. (2011) Structural and immunologic characterization of Ara h 1, a major peanut allergen. *J. Biol. Chem.*, **286**, 39318–39327.

17. ASTWOOD, J. D., LEACH, J. N. & FUCHS, R. L. (1996) Stability of food allergens to digestion in vitro. *Nat. Biotechnol.*, **14**, 1269–1273.
18. KOPPELMAN, S. J., HEFLE, S. L., TAYLOR, S. L. & DE JONG, G. A. H. (2010) Digestion of peanut allergens Ara h 1, Ara h 2, Ara h 3, and Ara h 6: a comparative in vitro study and partial characterization of digestion-resistant peptides. *Mol. Nutr. Food Res.*, **54**, 1711–1721.
19. BURKS, A. W., SHIN, D., COCKRELL, G., STANLEY, J. S., HELM, R. M. & BANNON, G. A. (1997) Mapping and mutational analysis of the IgE-binding epitopes on Ara h 1, a Legume Vicilin protein and a major allergen in peanut hypersensitivity. *Eur. J. Biochem.*, **245**, 334–339.
20. SHIN, D. S., COMPADRE, C. M., MALEKI, S. J., KOPPER, R. A., SAMPSON, H., HUANG, S. K., BURKS, A. W. & BANNON, G. A. (1998) Biochemical and structural analysis of the IgE binding sites on Ara h1, an abundant and highly allergenic peanut protein. *J. Biol. Chem.*, **273**, 13753–13759.
21. HOLM, L. & ROSENSTRÖM, P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.*, **38**(Suppl 2), W545–W549.
22. JÄGER, M., ZHANG, Y., BIESCHKE, J., NGUYEN, H., DENDLE, M., BOWMAN, M. E., NOEL, J. P., GRUEBELE, M. & KELLY, J. W. (2006) Structure–function–folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA*, **103**, 10648–10653.
23. ZAIDI, Z. H. & SMITH, D. L. (1996) *Protein Structure–Function Relationship*. Boston, MA: Springer.
24. UBERTO, R. & MOOMAW, E. W. (2013) Protein similarity networks reveal relationships among sequence, structure, and function within the cupin superfamily. *PLOS One*, **8**, 1–10.
25. LIU, H., MALHOTRA, R., PECZUH, M. W. & RUSLING, J. F. (2010) Electrochemical immunosensors for antibodies to peanut allergen Ara h2 using gold nanoparticle-peptide films. *Anal. Chem.*, **82**, 5865–5871.
26. NAKAMURA, R. & TESHIMA, R. (2013) Proteomics-based allergen analysis in plants. *J. Proteomics*, **93**, 40–49.
27. SHREFFLER, W. G. (2011) Microarrayed recombinant allergens for diagnostic testing. *J. Allergy Clin. Immunol.*, **127**, 843–849.
28. BRUSIC, V., PETROVSKY, N., GENDEL, S. M., MILLOT, M., GIGONZAC, O. & STELMAN, S. J. (2003) Computational tools for the study of allergens. *Allergy*, **58**, 1083–1092.
29. JIANG, B., QU, H., HU, Y., NI, T. & LIN, Z. (2007) Computational analysis of the relationship between allergenicity and digestibility of allergenic proteins in simulated gastric fluid. *BMC Bioinformatics*, **8**, 375.
30. NICOLAOU, N., MURRAY, C., BELGRAVE, D., POORAFSHAR, M., SIMPSON, A. & CUSTOVIC, A. (2011) Quantification of specific IgE to whole peanut extract and peanut components in prediction of peanut allergy. *J. Allergy Clin. Immunol.*, **127**, 684–685.
31. HILEMAN, R. E., SILVANOVICH, A., GOODMAN, R. E., RICE, E. A., HOLLESCHAK, G., ASTWOOD, J. D. & HEFLE, S. L. (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.*, **128**, 280–291.
32. LADICS, G. S., CRESSMAN, R. F., HEROUET-GUICHENEY, C., HERMAN, R. A., PRIVALLE, L., SONG, P., WARD, J. M. & MCCLAIN, S. (2011) Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. *Regul. Toxicol. Pharmacol.*, **60**, 46–53.
33. MCCLAIN, S. (2017) Bioinformatic screening and detection of allergen cross-reactive IgE-binding epitopes. *Mol. Nutr. Food Res.*, **61**, 1600676.
34. MOHABATKAR, H., MOHAMMAD BEIGI, M., ABDOLAH, K. & MOHSENZADEH, S. (2013) Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.*, **9**, 133–137.
35. SAHA, S. & RAGHAVA, G. P. S. (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.*, **34**, 202.
36. SCHEIN, C. H., IVANCIUC, O. & BRAUN, W. (2007) Bioinformatics approaches to classifying allergens and predicting cross-reactivity. *Immunol. Allergy Clin. North America*, **27**, 1–27.
37. DELVENNE, J.-C., YALIRAKI, S. N. & BARAHONA, M. (2010) Stability of graph communities across time scales. *Proc. Natl. Acad. Sci. USA*, **107**, 12755–12760.

38. DELMOTTE, A., TATE, E.W. AND YALIRAKI, S. & 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.
39. 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**, 1–11.
40. AMOR, B., YALIRAKI, S. N., WOSCHOLSKI, R. & BARAHONA, M. (2014) Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. *Mol. BioSyst.*, **10**, 2247–2258.
41. KORNEV, A. P. & TAYLOR, S. S. (2010) Defining the conserved internal architecture of a protein kinase. *Biochim. Biophys. Acta*, **1804**, 440–444.
42. LOCKLESS, S. W. & RANGANATHAN, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
43. SÜEL, G. M., LOCKLESS, S. W., WALL, M. A. & RANGANATHAN, R. (2002) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, **10**, 59–69.
44. XU, F., DU, P., SHEN, H., HU, H., WU, Q., XIE, J. & YU, L. (2009) Correlated mutation analysis on the catalytic domains of serine/threonine protein kinases. *PLoS One*, **4**, 1–11.
45. LAMBIOTTE, R., DELVENNE, J-C. & BARAHONA, M. (2014) Random walks, Markov processes and the multiscale modular organization of complex networks. *IEEE Trans. Netw. Sci. Eng.*, **1**, 76–90.
46. LANCICHINETTI, A. & FORTUNATO, S. (2009) Community detection algorithms: a comparative analysis. *Phys. Rev. E*, **80**, 056117.
47. NEWMAN, M. E. J. (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, **103**, 8577–8582.
48. BYRNE, S., ZHANG, H., MANN, D., BARAHONA, M. & YALIRAKI, S. (2017) Dynamic behaviour of the human cyclin-dependent kinases: a graph-theoretical analysis. Preprint.
49. DELMOTTE, A. (2015) All-scale structural analysis of biomolecules through dynamical graph partitioning. *Ph.D. Thesis*, Imperial College London.
50. WEBB, B. & SALI, A. (2002) *Comparative Protein Structure Modeling Using MODELLER*. Current Protocols in Bioinformatics, John Wiley & Sons, Inc., 5.6.1–5.6.32, 2014.
51. WORD, J., LOVELL, S. C., RICHARDSON, J. S. & RICHARDSON, D. C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
52. JACOBS, D. J., RADER, A. J., KUHN, L. A. & THORPE, M. F. (2001) Protein flexibility predictions using graph theory. *Proteins Struct. Funct. Genet.*, **44**, 150–165.
53. DAHIYAT, B. I., BENJAMIN GORDON, D. & MAYO, S. L. (1997) Automated design of the surface positions of protein helices. *Protein Sci.*, **6**, 1333–1337.
54. LIN, M. S., FAWZI, N. L. & HEAD-GORDON, T. (2007) Hydrophobic potential of mean force as a solvation function for protein structure prediction. *Structure*, **15**, 727–740.
55. DELVENNE, J-C., SCHAUB, M. T., YALIRAKI, S. N. & BARAHONA, M. (2013) The stability of a graph partition: a dynamics-based framework for community detection. *Dynamics On and Of Complex Networks* (A. Mukherjee, M. Choudhury, F. Peruani, N. Ganguly, B. Mitra eds), vol. 2. New York: Springer, pp. 221–242.
56. MELIGA, S. (2009) Graph clustering of atomic networks for protein dynamics. *Master's Thesis*, Imperial College London.
57. BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. & LEFEBVRE, E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, 10008–10019.
58. KARRER, B., LEVINA, E. & NEWMAN, M. E. J. (2008) Robustness of community structure in networks. *Phys. Rev. E*, **77**, 046119.
59. RONHOVDE, P. & NUSSINOV, Z. (2009) Multiresolution community detection for megascale networks by information-based replica correlations. *Phys. Rev. E*, **80**, 016109.
60. GOOD, B. H., DE MONTJOYE, Y-A. & CLAUSET, A. (2010) Performance of modularity maximization in practical contexts. *Phys. Rev. E*, **81**, 046106.

61. MEILÄ, M. (2003) Comparing clusterings by the variation of information. *Learning Theory and Kernel Machines* (B. Schölkopf, M. K. Warmuth eds), Berlin, Heidelberg: Springer, pp. 173–187.
62. MEILÄ, M. (2005) Comparing clusterings: an axiomatic view. *Proceedings of the 22nd International Conference on Machine Learning (ICML-05)* (S. Dzeroski, L. De Raedt, S. Wrobel eds), vol. 1. Bonn, Germany: ACM Press (US), pp. 577–584.
63. MEILÄ, M. (2007) Comparing clusterings—an information based distance. *J. Multivariate Anal.*, **5**, 873–895.
64. RASMUSSEN, C. E. & WILLIAMS, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts, London, England: The MIT Press.
65. JUST, V. J., STEVENSON, C. E. M., BOWATER, L., TANNER, A., LAWSON, D. M. & BORNEMANN, S. (2004) A closed conformation of *Bacillus subtilis* oxalate decarboxylase OxdC provides evidence for the true identity of the active site. *J. Biol. Chem.*, **279**, 19867–19874.