Multi-Scale Graph Partitioning for Unravelling Dynamics of Major Peanut Allergen Ara h 1

Heng Zhang
Advanced Software Technology Innovation Lab
OnePlus Technology Co. Ltd.
andrew.zhangh@oneplus.com

1 Introduction

Proteins are multi-scale biomolecular machines with coupled structural organizations across time and length scales. Each level of these structural organizations links to some functional behaviour and the related scales can span across several orders of magnitude [1]. Understanding protein dynamics across scales is especially important in a problem such as peanut allergy, an immunoglobulin E (IgE)-mediated hypersensitivity [10]. The mechanism of an allergy reaction elicitation is so poorly understood that there is yet no methodology that can *a priori* robustly predict the allergenicity of a protein.

Allergen protein studies has been focused on epitopes and critical residues. Epitopes are linear fragments of digested proteins that can bind to IgE in experiments. The major peanut allergen Ara h 1 is identified with at least 23 stable epitopes [13, 11, 12]. Among these epitopes, No. 8 and 14 are immunodominant, as they can bind to most of the IgE samples. Site-directed mutagenesis showed that point mutations at certain residues have strong impact on their binding affinity [14]. Therefore, these residues are called *critical residues*. Negative proteins are structurally similar to allergens, but they do not exhibit allergenic potential. Distinguishing between allergens and negative controls remains difficult computationally. Oxalate decarboxylase (OxdC) shares 40% sequence similarity and 68.6% structural similarity with Ara h 1. Another negative control, MnCA (Mn2+-cupin A), is 21% sequentially and over 68.4% structurally similar to Ara h 1. Neither has been shown to trigger allergenicity so far, even though they are structurally clustered together within their superfamily [15]. 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 series of computational methods have been developed to extract information from allergen proteins, and infer their allergenic potential. However, differentiating similar proteins based on their allergy-triggering capability is still difficult [16, 17, 18, 19]. If not large-scale molecular simulation studies that is specifically designed and optimized, atomic-level simulation methods are still prevented from reaching functionality related scales [2, 3, 4, 5, 6]. In order to reach the scales where global dynamics of proteins are accessible, coarse-graining methods have been proposed with certain levels of success [7, 8, 9], at a cost of loss of generality and of smearing the detailed physico-chemical atomic interactions.

We address the question of peanut allergenicity from a dynamics perspective, by studying the dynamics on graphs that are encoded with protein structural information at the atomistic level. In the meantime, we develop a computational mutation method to identify *hot spots*, residues that impact the global dynamics of an allergen. So far as we know, this is the first time that a multi-scale graph-partitioning method is applied for allergy understanding.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

2 Methodology

We apply Markov Stability [21] to analyse the allergen graph, an undirected, weighted atomic graph generated from the allergen protein conformation using energy functions. The conformation can be obtained from the RCSB Protein Data Bank [20], amongst many public structural repositories. By optimizing Markov Stability value at each Markov time, we find the optimized partition where a random walk is most likely to remain. Significant partitions are obtained in increasingly coarser communities. In the case of a protein graph, this process allows us to scan across resolutions and find clusters corresponding to meaningful biochemical groups at different granularities. These groupings correspond to groups of atoms moving coherently across certain scales. 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 and weighted graph $G(V, E)$, denoted by the adjacency matrix $\mathbf{A}$ of rank n. The vertex degrees of the graph are $d_i = \sum_{ij}^{n} A_{ij}$, and the degree matrix is defined as $\mathbf{D} = diag(\mathbf{d})$. On such a graph, we consider a continuous-time Markovian process that is governed by the combinatorial Laplacian, $\mathbf{L} = \mathbf{D} - \mathbf{A}$, as most appropriate for protein dynamics:

$$\frac{d\mathbb{P}}{dt} = -\frac{1}{\langle d \rangle}\mathbf{pL} \tag{1}$$

where $\langle d \rangle = (\mathbf{1}^T D \mathbf{1})/N$ is the average degree and $\mathbb{P} = \mathbf{pD}^{-1}$. Now the stationary distribution correspondingly is the uniform distribution over all nodes: $\pi = \mathbf{1}^T/N$.

Markov Stability is then defined as

$$r(t, P) = tr(\mathbf{H}^T[\Pi_c e^{-\mathbf{L}t/\langle \mathbf{d} \rangle} - \pi_c^T \pi_c]\mathbf{H}) \tag{2}$$

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 1 if node i belongs to community j and 0 otherwise; π_c denotes the stationary distribution defined above, and Π_c are the diagonal elements of π_c. The time t is the Markov time or a dimensionless resolution parameter.

We use the Louvain algorithm, a greedy agglomerative method [22], to solve the partitioning problem. The Louvain algorithm is deterministic, but the final solution depends on the order in which the different nodes are scanned initially. This initial ordering, or the Louvain initial condition, can be chosen at random every time the calculation is executed. We use the variability induced by our random choice of the Louvain initial conditions, i.e. the Variation of Information (VI) [23], to estimate the robustness of a partition. Other perturbations affecting edge weights or the quality function, for example, have been considered in the past and shown to yield similar results [21]. We also developed a linearised version of Markov Stability, which can produce similar partition results, with 20X times or more speed-up, depending on the protein structure.

Another question of interest is the identification of *hot spots* that significantly impact the protein global dynamics when altered locally. To mimic *in silico* the process of this mutation procedure on the protein graph, we remove all edges representing weak interactions with respect to the side chain of an amino acid node group. Then, the mutated graph is partitioned using Markov Stability. We identify the mutations by filtering the outliers of robustness ensemble of multiple mutations along with the VI vector of the original protein graph, using Gaussian process regression (GPR) [24]. The calculation can be realized using public libraries such as the gpml MATLAB toolbox (http://www.gaussianprocess.org/gpml/code/).

3 Experiments

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 (Figure 1). From Markov time of 105 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

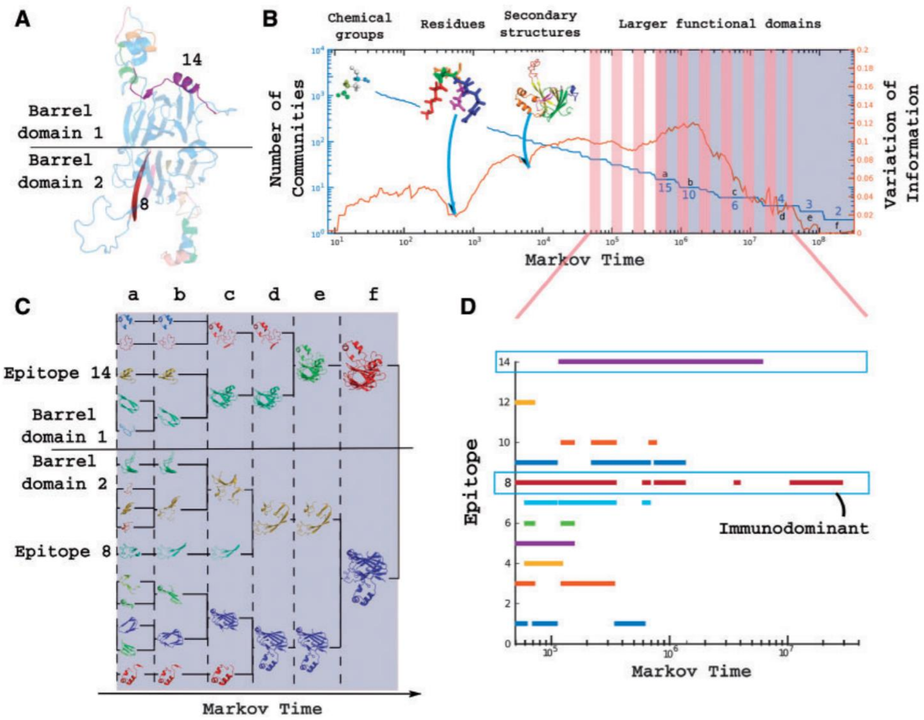

Figure 1: Structural anatomy of the allergen Ara h 1 at all scales. (A) The atomistic structure of the core of the monomer of Ara h 1 (PDB ID: 3S7E), (B) Markov Stability analysis of the core of the Ara h 1 monomer. As the Markov time increases, we recover first the meaningful biochemical levels of organization (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.

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 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 epitope 14 is over twice the size of others, its failing threshold was set to 30%. No community can be mapped with epitopes at larger scales, because smaller communities will merge into large functional domains, indicating global dynamics. Epitopes 8 and 14 last much longer than the others, whereas epitopes 2, 11 and 13 did not manifest themselves as single communities. Epitope 8 is an immunodominant epitope, 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.

We compare the Markov Stability analysis results of Ara h 1 and its two controls. As discussed in the Introduction section, the three proteins are structurally similar and share the two-barrelled configuration. 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.

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.

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 h 1 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. [25] argued that E162 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.

We finally show 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 E222 and H211. E222 is located by epitope 5, whereas H211 is within epitope 14, beside the partitioning boundary residue A212 at medium scales, and not far from E222. 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.

4 Conclusion

We studied the major peanut allergen Ara h 1 as well as its structurally similar negative controls through an atomistic-based graph partitioning methodology, based on Markov Stability, as a viable approach for unravelling protein dynamics across multiple scales. By partitioning the graphs generated from the 3D protein conformation, we are able to find meaningful communities at different resolutions related to scales 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.

References

[1] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964, 2007.

[2] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646, 2002.

[3] Martin Karplus and John Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences*, 102(19):6679–6685, 2005.

[4] John L Klepeis, Kresten Lindorff-Larsen, Ron O Dror, and David E Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Current opinion in structural biology*, 19(2):120–127, 2009.

[5] Ambuj Kumar and Rituraj Purohit. Use of long term molecular dynamics simulation in predicting cancer associated snps. *PLoS computational biology*, 10(4):e1003318, 2014.

[6] Danijela Apostolovic, Dragana Stanic-Vucinic, Harmen HJ De Jongh, Govardus AH De Jong, Jelena Mihailovic, Jelena Radosavljevic, Milica Radibratovic, Julie A Nordlee, Joseph L Baumert, Milos Milcic, et al. Conformational stability of digestion-resistant peptides of peanut conglutins reveals the molecular basis of their allergenicity. *Scientific reports*, 6:29249, 2016.

[7] Philippe Derreumaux and Normand Mousseau. Coarse-grained protein molecular dynamics simulations. *The Journal of chemical physics*, 126(2):01B608, 2007.

[8] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David Van Der Spoel, et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[9] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical reviews*, 116(14):7898–7936, 2016.

[10] Xiumei Hong, Ke Hao, Christine Ladd-Acosta, Kasper D Hansen, Hui-Ju Tsai, Xin Liu, Xin Xu, Timothy A Thornton, Deanna Caruso, Corinne A Keet, et al. Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in us children. *Nature communications*, 6:6304, 2015.

[11] James D Astwood, John N Leach, and Roy L Fuchs. Stability of food allergens to digestion in vitro. *Nature biotechnology*, 14(10):1269, 1996.

[12] Stef J Koppelman, Sue L Hefle, Steve L Taylor, and Govardus AH De Jong. 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. *Molecular nutrition & food research*, 54(12):1711–1721, 2010.

[13] A Wesley Burks, David Shin, Gael Cockrell, J Steven Stanley, Ricki M Helm, and Gary A Bannon. Mapping and mutational analysis of the ige-binding epitopes on ara h 1, a legume vicilin protein and a major allergen in peanut hypersensitivity. *European Journal of Biochemistry*, 245(2):334–339, 1997.

[14] David S Shin, Cesar M Compadre, Soheila J Maleki, Randall A Kopper, Hugh Sampson, Shau K Huang, A Wesley Burks, and Gary A Bannon. Biochemical and structural analysis of the ige binding sites on ara h1, an abundant and highly allergenic peanut protein. *Journal of Biological Chemistry*, 273(22):13753–13759, 1998.

[15] Richard Uberto and Ellen W Moomaw. Protein similarity networks reveal relationships among sequence, structure, and function within the cupin superfamily. *PLoS One*, 8(9):e74477, 2013.

[16] Ronald E Hileman, Andre Silvanovich, Richard E Goodman, Elena A Rice, Gyula Holleschak, James D Astwood, and Susan L Hefle. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *International archives of allergy and immunology*, 128(4):280–291, 2002.

[17] V Brusic, N Petrovsky, SM Gendel, M Millot, O Gigonzac, and SJ Stelman. Computational tools for the study of allergens. *Allergy*, 58(11):1083–1092, 2003.

[18] Bingjun Jiang, Hong Qu, Yuanlei Hu, Ting Ni, and Zhongping Lin. Computational analysis of the relationship between allergenicity and digestibility of allergenic proteins in simulated gastric fluid. *BMC bioinformatics*, 8(1):375, 2007.

[19] Scott McClain. Bioinformatic screening and detection of allergen cross-reactive ige-binding epitopes. *Molecular nutrition & food research*, 61(8):1600676, 2017.

[20] Peter W Rose, Andreas Prlić, Ali Altunkaya, Chunxiao Bi, Anthony R Bradley, Cole H Christie, Luigi Di Costanzo, Jose M Duarte, Shuchismita Dutta, Zukang Feng, et al. The rcsb protein data bank: integrative view of protein, gene and 3d structural information. *Nucleic acids research*, page gkw1000, 2016.

[21] J-C Delvenne, Sophia N Yaliraki, and Mauricio Barahona. Stability of graph communities across time scales. *Proceedings of the national academy of sciences*, 107(29):12755–12760, 2010.

[22] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[23] Marina Meilă. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer, 2003.

[24] CE Rasmussen and CKI Williams. Gaussian processes for machine learning. the mit press, cambridge, massachusetts, usa, london, 2006.

[25] Victoria J Just, Clare EM Stevenson, Laura Bowater, Adam Tanner, David M Lawson, and Stephen Bornemann. A closed conformation of bacillus subtilis oxalate decarboxylase oxdc provides evidence for the true identity of the active site. *Journal of Biological Chemistry*, 279(19):19867–19874, 2004.