VERMONT: Visualizing mutations and their effects on protein physicochemical and topological property conservation

Sabrina A. Silveira* Valdete M. Goncalves-Almeida* Yussif Barcelos* Elisa B. de Lima*
Flavia F. Aburjaile Laerte M. Rodrigues Wagner Meira Jr. Raquel C. de Melo-Minardi*

*These authors contributed equally to this work
Universidade Federal de Minas Gerais, Brazil

ABSTRACT

In this paper, we propose an interactive visualization called VERMONT (ViewER MutatIOn Tool), which tackles the problem of visualizing mutations and their effects on the conservation of physicochemical and topological properties in protein families. More specifically, we visualize multiple sequence alignments based on structural alignments and integrate several structural parameters that should aid biologists in gaining insights on possible mutation consequences. With this tool, we were able to identify patterns of position-specific properties as well as exceptions that could help predict whether specific mutations could damage protein function.

Availability: www.dcc.ufmg.br/~sabrinas/vermont/

1 INTRODUCTION

Some DNA mutations (i.e., substitutions, insertions and deletions), which occur naturally due to evolutionary pressure, are known to affect protein function. A significant open problem in science and in Bioinformatics consists of collecting, integrating and processing a huge amount of information, and ultimately presenting it in a simple and visual manner in order to facilitate both comprehension by scientists and the solution of intriguing problems. How may specific modifications in protein amino acid properties help to identify potentially significant mutations? Depending on where these mutations occur, a protein can lose its function or become inactive. Some mutations can cause destabilization and significant structural modifications. The goal of this work is to develop a visualization tool capable of identifying mutations and pointing out possible consequences for protein function.

2 METHODS

Although sequence-based predictors have shown good performance, prediction quality can be further improved by introducing features derived from the three-dimensional protein structures [3]. In this work, we consider sequence conservation from a structural perspective by performing structure-based sequence alignments. Also, we enrich the data set with several structural features such as solvent accessibility, presence in a structural cavity or at the active site, as well as various metrics computed from the chemical interactions that a residue can establish with its structural neighbors.

2.1 Data Set

Raw sequences and structures

We obtained the protein sequences from the BioVis Contest Organization and searched for each of them in the Protein Data Bank (PDB). We then retrieved every structure corresponding to one of the provided TIM sequences, resulting in the set of 133 enzymes used in this work. All sequences were aligned using Multiprot [16].

The structures on PDB files were improved using the PDBEST package [11], a tool developed by our group that accesses and evaluates the annotation quality PDB files. In this work, we applied a set of rules to improve our data set, including, for instance, chain separation and removal of identical chains or structures, as well as of chains with missing atoms. In case of structure models, we use the first one in the PDB file.

Solvent Accessibility

Solvent accessibilities were computed by software NAccess1, which implements the algorithm developed by Lee and Richards [7]: a probe of a given radius is rolled around the surface of the protein, and the accessible surface is defined as the path traced by the probe’s center. Typically, the 1.4 Å radius of a water molecule is used, thus the surface is described as the solvent-accessible surface. Absolute accessibility is given in Å². Since each amino acid presents distinct volume and surface area, we preferred to work with relative accessibilities, which express the accessible surface as a percentage of that observed in a ALA-X-ALA tripeptide (to mimic the extended conformation).

Active Sites

We retrieved active site information from Catalytic Site Atlas (CSA) [13], a database documenting enzyme active sites and catalytic residues. CSA defines active site residues either based on bibliographic references or, more automatically, on remote homology computed using PSI-BLAST alignments.

Chemical Interactions

In previous works [5, 12, 6], we have shown successful approaches for studying protein functional signatures through the precise computation and detection of chemical interaction patterns between protein amino acids. In this work, we use a cutoff-independent approach to geometrically compute probable amino acid interactions at atomic level, which we mapped to residue level. For each protein, we generate a Voronoi diagram followed by its Delaunay tesselation [14, 10] and, using both distance and physicochemical properties described in [17], we classify contacts into one of the following types: charged attractive, charged repulsive, aromatic, hydrophobic and hydrogen bonds. In this process we use the CGAL software library [1] and Gromacs [18].

Topological Properties

In [8], authors used complex networks to study the role of an amino acid in both local and global structures, as well as to determine the extent to which disease-associated mutations and Single Amino Acid Polymorphisms (SAPs) differ in terms of their interactions with other residues. They showed that mutations are probably disease-associated when they occur at a site with a high centrality value and/or high degree value in the network. They also showed

1http://www.bioinf.manchester.ac.uk/naccess/
that neighboring residues around a mutation site can help to determine whether the mutation is associated with diseases or not.

The complex network measures used in this work, which are computed using the igraph package [4] from R software [15], are defined as follows.

- **The degree** \( k_i \) of vertex \( i \) in a graph is the number of edges connected to it [9]. For an undirected graph containing \( n \) vertices, the degree can be written in terms of the adjacency matrix as \( k_i = \sum_{j=1}^{n} A_{ij} \).

- **Betweenness** is a centrality measure that represents the extent to which a vertex lies on paths among other vertices. Mathematically, let \( n'_{ij} \) be 1 if vertex \( i \) lies on the geodesic path from \( s \) to \( t \), and 0 if it doesn’t or if there is no such path (because \( s \) and \( t \) lie in different network components). Then, the betweenness centrality \( x_i \) is given by \( x_i = \sum_{s,t \neq i} n'_{ij} \). This definition considers separately the geodesic paths in either direction between each vertex pair. Since our network is undirected, this effectively counts each pair twice, which is compensated by dividing the result by 2. Vertices with high betweenness centrality may have considerable influence within a network due to their control over information passing among others.

- **The closeness** centrality measure represents the average distance from a vertex to all other vertices. Suppose \( d_{ij} \) is the length of a geodesic path from \( i \) to \( j \) (i.e., the number of edges along the path). Then, the mean geodesic distance from \( i \) to \( j \), averaged over all vertices \( j \) in the network, is \( l_i = \frac{1}{n-1} \sum_j d_{ij} \). This measure takes low values for vertices that are separated from others by only a short geodesic distance on average. Such vertices might have better access to information at other vertices or more direct influence on them.

### 2.2 Visualizations

In this section, we describe the proposed visualizations, discuss some of the challenges posed by the data we had and the requirements that visualizations were expected to meet, as well as some project decisions we have made. All visual representations were developed in D3 [2].

The data set was originally composed of a huge set of sequences with about 250 positions, each one representing a biological object: a specific amino acid residue. We computed several physicochemical and topological properties for each of these residues, which demanded a visualization capable of representing these data as well as the sequences. Therefore, we have a multivariate quantitative visualization problem which requires a tool that allows domain specialists to compare multiple sequences and their parameters revealing patterns and exceptions.

It is well known that random insertions, deletions and substitutions on the nucleotide sequence within a gene may change the amino acid sequence of the corresponding protein. Some of these mutations do not drastically alter the protein’s structure, but others do, thus impairing the protein’s ability to function. Therefore, alignment strategies must be able to properly compare protein sequences. Once such robust alignments are available, proper visualization techniques are required to make sense of the similarities and dissimilarities between the set of sequences.

Biologists are used to visualizing and analyzing sequence alignments, with classical visualizations consisting of depicting each sequence on a row, with columns representing equivalent positions in the alignment. It is common to use color codes to aid the spotting of relevant conservations and exceptions in columns. Thus, an important requirement of the present visualization challenge was to stick to these traditional visual representations very easily interpreted by biologists. Consequently, the traditional visualization of multiple sequence alignments is the basis for those produced in this work and is shown in Figure 1. In addition to displaying aligned sequences, we include a varied set of physicochemical and topological parameters and many techniques for interaction analysis.

![Figure 1: Basis for the proposed visualizations with CINENA color scheme. Rows represent sequences and columns represent equivalent positions in the structure-based sequence alignment.](Adapted from www.bioinformatics.nl/berndb/aacolour.html)
Physicochemical and Topological Properties

This visualization combines the traditional sequence alignment visualization with a heatmap. It shows, using color intensity, the following measures: relative solvent accessibility, degree, betweenness and closeness. It helps users to spot conserved properties at specific alignment positions.

On all panels, users may obtain details on demand by passing the mouse over a residue cell. The tool will show residue type, position in the structure-based alignment and real position in the original sequence, as well as the values for all the computed parameters.

3 RESULTS AND DISCUSSIONS

As previously mentioned, we decided to work only with sequences whose 3D structures were available, since structures are much more conserved than sequences. In other words, a set of very dissimilar sequences can fold in a very similar way preserving an identical function. Based on the alignment and analysis of the mutant sequence (dTIM) and its wild type protein (PDB ID 2YPI:A), we identified 96 mutations to study. Each single point mutation was studied from different perspectives: family residue conservation as well as physicochemical and topological properties.

This work's main hypothesis is that important residues are conserved through evolution, so conserved positions in the multiple sequence alignment are important residues for function preservation. As previously discussed, some mutations do not drastically alter a protein’s structure and function, but others do. Some amino acid substitutions are commonly found throughout the molecular evolution process, while others are rare: Asparagine, Aspartate, Glutamate and Serine are the most mutable amino acids, while Cysteine and Tryptophan are the least. Besides, it is important to highlight that the substitution of an amino acid by another with similar physicochemical properties will probably not influence protein stability. Having this in mind, we classified amino acids according to three different schemes (CINEMA, CLUSTAL and LESK) and analyzed mutations from a conservative point of view.

We prioritized non-conservative mutations as they are probably have larger impact on protein function. For each mutation, we verified if it was frequent, rare or very rare in the family. Frequent mutations probably do not impact function, since they appeared in other proteins of the family at a similar context and did not imply in function loss. We decided to further investigate only the rare and very rare mutations. One of our first assessments when analyzing the contest data was that every mutation was present in at least one
other sequence of the family, which meant that no trivial mutations causing loss of function were found.

It is also important to mention that no mutations were found in protein active sites (N10, K12, H95, E165 and G171). Such mutations would obviously impact function. We have also investigated possible mutations in residues that are in contact with the active site, which would probably affect active site conformation and lead to function modification. However, no such mutation was found.

The non-conservative rare and very rare mutations were then carefully investigated considering accessibility and topological properties, leading to the set of significant mutations presented on Table 1. Underlined mutations have higher probability of causing function loss.

Table 1: Predictions of deleterious mutations: 22 mutations identified as possibly causing damage to protein function. Underlined mutations are predicted with more confidence according to the evidences used. Sequence numbering according to PDB ID 2YPI:A

| Mutation | Avg. degree | Avg. betweenness | Avg. closeness (E-04) | Avg. accessibility |
|----------|-------------|------------------|-----------------------|-------------------|
| S19E     | 3.93        | 101.68           | 9.9                   | 47.47             |
| I20A     | 8.45        | 728.61           | 9.9                   | 3.12              |
| N28K     | 5.54        | 93.93            | 9.9                   | 30.09             |
| K55G     | 6.87        | 198.56           | 9.9                   | 16.32             |
| T60K     | 5.25        | 455.12           | 9.8                   | 25.92             |
| K69E     | 4.57        | 362.98           | 10.0                  | 39.75             |
| S71K     | 2.86        | 160.55           | 9.7                   | 69.87             |
| K89D     | 3.82        | 289.67           | 9.9                   | 44.37             |
| D111K    | 5.08        | 89.62            | 9.8                   | 44.26             |
| G118E    | 3.53        | 19.84            | 9.8                   | 87.12             |
| E152A    | 3.80        | 22.26            | 9.9                   | 73.91             |
| E153G    | 5.91        | 120.75           | 9.9                   | 40.56             |
| K155D    | 3.49        | 23.58            | 9.8                   | 74.42             |
| T158K    | 4.39        | 100.68           | 9.9                   | 60.49             |
| S194E    | 4.37        | 26.71            | 9.9                   | 74.15             |
| K195N    | 4.61        | 49.99            | 9.9                   | 55.20             |
| K199E    | 3.56        | 56.41            | 9.9                   | 64.94             |
| S202E    | 4.35        | 100.48           | 9.9                   | 52.74             |
| N213K    | 5.72        | 265.96           | 9.9                   | 37.79             |
| G214P    | 4.09        | 108.98           | 9.9                   | 31.33             |
| K221A    | 5.24        | 246.48           | 10.0                  | 19.78             |
| D222A    | 4.89        | 56.55            | 9.9                   | 68.43             |

4 CONCLUSION

In this paper, we propose VERMONT, an interactive tool to visualize mutations and their possible consequences for protein structure and function. We modeled the problem as spotting residue conservations together with the conservation of physicochemical and topological properties. The proposed interactive visualization provides a macro view of the multiple structure-based sequence alignment as well as several other structural parameters.

The tool allows users to view, at a glance, a multivariate set of residue parameters, expanding or compressing panels and zooming out to see full-length of sequences, or zooming in to focus on parts of them. Users may also filter residues individually or in groups of similar properties by highlighting or attenuating them, which helps in spotting patterns and exceptions.

Using the proposed visualization tool, we were able to predict 22 mutations we believe have a significant probability of causing damage to protein function, four of which seem to be more severe and have a high likelihood of causing function loss.

ACKNOWLEDGEMENTS

This work was supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Financiadora de Estudos e Projetos (FINEP) and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

REFERENCES

[1] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.
[2] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. Visualization and Computer Graphics, IEEE Transactions on, 17(12):2301–2309, 2011.
[3] E. Capriotti and R. Altman. Improving the prediction of disease-related variants using protein three-dimensiona structure. BMC Bioinformatics, 12(Suppl 4):S3, 2011.
[4] G. Csardi and T. Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems:1695, 2006.
[5] C. da Silveira, D. Pires, R. Melo-Minardi, C. Ribeiro, C. Veloso, J. Lopes, W. M. Jr., G. Neshich, C. Ramos, R. Habesch, and M. Santoro. Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. Proteins, 74(3):727–743, 2009.
[6] V. Gonçalves-Almeida, D. E. Pires, R. C. de Melo-Minardi, C. H. da Silveira, W. Meira, and M. M. Santoro. Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. Bioinformatics, 28(3):342–349, 2012.
[7] B. Lee and F. Richards. The interpretation of protein structures: estimation of static accessibility. Journal of Molecular Biology, 55(3):379–400, 1971.
[8] Y. Li, Z. Wen, J. Xiao, H. Yin, L. Yu, and M. Li. Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. BMC Bioinformatics, 12(14), 2011.
[9] M. Newman. Networks. Oxford University Press, 1st edition edition, 2010.
[10] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. Spatial tessellations: concepts and applications of Voronoi diagrams, volume 501. Wiley, 2009.
[11] D. Pires, C. da Silveira, M. Santoro, and W. P. Meira Jr. Pdb enhanced structures toolkit. In Proceedings of the 3rd International Conference of Brazilian Association for Bioinformatics and Computational Biology, 2007.
[12] D. Pires, R. Melo-Minardi, M. dos Santos, C. da Silveira, M. Santoro, and W. M. Jr. Cutoff scanning matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC Genomics, 12(Suppl 4):S12, 2011.
[13] C. T. Porter, G. J. Bartlett, and J. M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Research, 32:D129–D133, 2004.
[14] A. Poupon. Voronoi and voronoi-related tessellations in studies of protein structure and interaction. Current opinion in structural biology, 14(2):233–241, 2004.
[15] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2011.
[16] M. Shatsky, R. Nussinov, and H. H. Wolfson. A method for simultaneous alignment of multiple protein structures. Proteins: Structure, Function and Bioinformatics, 56(1):143–156, 2004.
[17] V. Sobolev, A. Sorokine, J. Prilusky, E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. Bioinformatics, 15(4):327–332, 1999.
[18] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. Gromacs: fast, flexible, and free. Journal of computational chemistry, 26(16):1701–1718, 2005.