bioRxiv preprint doi: https://doi.org/10.1101/520833; this version posted April 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GAPIN: Grouped and Aligned Protein Interface Networks

Biharck M. Araújo[1], Aline L. Coelho[2], Sabrina A. Silveira[1,3,4], João P. R. Romanelli[5], Raquel C. de Melo-Minardi[1,6] and Carlos H. da Silveira[1,7]

[1] Bioinformatics PhD Program, Universidade Federal de Minas Gerais, 31270-901, Brazil.

[2] School of Medicine - EMED / Universidade Federal de Ouro Preto, 35400-000, Brazil.

[3] Department of Computer Science, Universidade Federal de Viçosa, 36570-900, Brazil.

[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK.

[5] Institute of Applied and Pure Sciences, Universidade Federal de Itajubá, 35903-087, Brazil.

[6] Department of Computer Science, Universidade Federal de Minas Gerais, 31270-901, Brazil.

[7] Institute of Technological Sciences, Universidade Federal de Itajubá, 35903-087, Brazil.

Abstract

Summary: GAPIN is a web-based application for structural interaction network analysis among any type of PDB molecules, regardless of whether their interfaces are between chain-chain or chain-ligand. A special emphasis is given to graph clustering, allowing users to scrutinize target contexts for ligand candidates. We show how GAPIN can be used to unveil underlying hydrophobic patterns on a set of peptidase-inhibitor complexes. In another experiment, we show there is a positive correlation between cluster sizes and the presence of druggable spots, indicating that the clustering may discriminate the higher complexity of these hot subnetworks.

Availability and implementation: GAPIN is freely available as an easy-to-use web interface at https://gapin.unifei.edu.br.

Contact: biharck@ufmg.br, carlos.silveira@unifei.edu.br.

Supplementary information: Supplementary data are available.

Introduction

Several computational methods, databases, and original tools have been proposed to help analyze the profusion of biomolecular interaction data generated by experimentalists (Cafarelli et al., 2017), hoping to extract useful knowledge mainly through comprehensive mappings (Barabási et al., 2011), such as NetworkAnalyzer/RINalyzer (Doncheva et al.,2012), STRING (Szklarczyk et al, 2017), Bio3D-web (Skjærven et al., 2016), NAPS (Chakrabarty et al.,2016), STITCH (Szklarczyk et al., 2016), among others (McGillivray et al.,2018). However, these initiatives usually have their own focus (on protein-protein interfaces or protein-chemical interactions, for instance) and they do not easily integrate different types of interfaces, such as

bioRxiv preprint doi: https://doi.org/10.1101/520833; this version posted April 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chain-chain or chain-ligand, regardless of the biomolecule involved. Moreover, how can we best visualize these interaction networks with the design of new drug candidates in mind?

To join efforts to this challenge, we propose GAPIN - *Grouped and Aligned Protein Interface Networks*, a fully web-based application with immediate availability, portability, and all friendly convenience that a modern browser can offer. GAPIN has RCSB PDB (Berman et al., 2000) coordinates as main input, defining interfaces among biomolecules as graphs with node representations at the atomic level. This permits visualizations and analysis regardless of whether the interfaces are among proteins, nucleic acids, carbohydrates, lipids, chemical ligands, ions or even waters. GAPIN contrasts the PDB structures and correspondent graph models with different view options. It introduces a network modularity analysis, by clustering the atomic graphs into communities of densely connected nodes, forming higher-level or modular graphs. GAPIN provides a graph-structure alignment tool for comparative interface analysis. Higher-level graphs also help identify spots, a group of a few residues in chain-chain interactions that can have a relevant contribution to free energy of binding. Consequently, they can be candidates for new druggable targets (Xia et al., 2010). To highlight the wide scope of GAPIN, we perform two experiments: one on graph alignments in a peptidase-inhibitor data set, and another on spot locations in an alanine mutagenesis data set.

Material and Methods

The data set for the interface alignment experiment was taken from (Gonçalves-Almeida et al., 2012), and for spot locations from (Xia et al., 2010). In lower-level graphs, nodes are atoms, and edges are weighted by contact areas. Higher-level graphs are built by a spectral graph clustering described in (Luxburg, 2007). For more details, see Supplementary Material.

Interface Alignments

Gonçalves-Almeida et al. (2012) showed how hydrophobic patterns in peptidase-inhibitor interfaces can help explain the cross-inhibition phenomenon, when dissimilar structural chains bind to a particular site promiscuously. Classical examples are some trypsin-like and subtilisin-like serine-peptidases. Notwithstanding their divergent 3D structures (all beta against alpha and beta) and sequence identity as low as 20%, they can be cross-inhibited by both Turkey ovomucoid or Eglin C, also two different inhibitors. The authors indicated the existence of conserved hydrophobic patches on such interfaces using a nonredundant set of 9 peptidase-inhibitor complexes. (see Table S2). Figure 1A shows the result of chain-chain interface alignments for 2 of these complexes, comprising nonpolar-nonpolar contacts in 1PPF

bioRxiv preprint doi: https://doi.org/10.1101/520833; this version posted April 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Leukocyte elastase - Turkey ovomucoid 3) and 1TEC (Thermitase subtilisin - Eglin C). It is possible to see the graph similarities and the notable superimposition of nodes, in agreement with aforementioned results, but with the differential that GAPIN performs it for any PDB chain-chain interfaces (chain-ligand interfaces soon). See Figure S2 and S3 for the remaining alignments and RMSD statistics, respectively.

Spot Locations

At chain-chain interfaces, alanine mutations are the common procedure used to discriminate residues that account for most of the free energy of binding (Xia et al., 2010). If the impact of this mutation causes a change in ΔG of binding at least 2.0 $kcal.mol^{-1}$, then the target residue is called a hot spot. If less than this, a warm spot. If greater than 4.0 $kcal.mol^{-1}$, a red-hot spot. Xia et al. compiled an experimentally determined spot data set comprising 15 protein-protein complexes (see “Additional file 1” in this citation). We mapped the presence of these spots into the higher graphs, according to the node location. Node size distribution plots can be seen in Figure 1B for red-hot spots (red/solid), hot spot (orange/dashed) and warm spots (purple/dotdash) against the others (blue/dotted). The difference between the distributions is statically significant (at confidence level of 0.95) for red-hot and hot spots (KS-test with p-value 0.012 and 0.019, respectively) but not for warm spots (p-value 0.32). This indicates that larger nodes may be more likely to host druggable spots. GAPIN allows users to isolate and explore such nodes in different ways (Figure S4). See Table S3 for more statistical evidence.

Conclusion

GAPIN is a modern web-based application focused on structural interaction network analysis of PDB biomolecule interfaces. Herein, we show how it can be used to unveil hydrophobic patterns between different peptidase-inhibitor complexes. We also show there is a positive correlation between node sizes and the presence of energetic spots. Probably because these interface regions involve more complex shape complementarity and denser interactions, resulting in more intricate subnetworks, which is in line with other studies (Moreira et. al.,2007).

Acknowledgements

We are grateful to Prof. Ricardo Leão/USP by the inspiration of the name GAPIN.

bioRxiv preprint doi: https://doi.org/10.1101/520833; this version posted April 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Funding

This work has been supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [477587/2013-5 to C.H.S.];

Conflict of Interest: none declared.

References

Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature Reviews. Genetics, 12(1), 56–68.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. Nucleic Acids Research, 28(1), 235–242.

Cafarelli, T. M., Desbuleux, A., Wang, Y., Choi, S. G., De Ridder, D., & Vidal, M. (2017). Mapping, modeling, and characterization of protein-protein interactions on a proteomic scale. Current Opinion in Structural Biology, 44, 201–210.

Chakrabarty, B., & Parekh, N. (2016). NAPS: Network analysis of protein structures. Nucleic Acids Research, 44(W1)

Doncheva, N. T., Assenov, Y., Domingues, F. S., & Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. Nature Protocols, 7(4), 670–685.

Gonçalves-Almeida, V. M., Pires, D. E. V, de Melo-Minardi, R. C., da Silveira, C. H., Meira, W., & Santoro, M. M. (2012). HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. Bioinformatics (Oxford, England), 28(3), 342–9.

Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing, 17(4), 395–416.

McGillivray, P., Clarke, D., Meyerson, W., Zhang, J., Lee, D., Gu, M., … Gerstein, M. (2018). Network Analysis as a Grand Unifier in Biomedical Data Science. Annual Review of Biomedical Data Science, 1(1), 153–180.

Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Hot spots—A review of the protein–protein interface determinant amino-acid residues. Proteins: Structure, Function, and Bioinformatics, 68, 803–812

Skjærven, L., Jariwala, S., Yao, X. Q., & Grant, B. J. (2016). Online interactive analysis of protein structure ensembles with Bio3D-web. Bioinformatics, 32(22), 3510–3512.

Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Research, 44(D1), D380–D384.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., … Von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Research, 45(D1), D362–D368.

Xia JF, Zhao XM, Song J, & Huang DS. (2010). APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics, 11, 174.

bioRxiv preprint doi: https://doi.org/10.1101/520833; this version posted April 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

List of figures

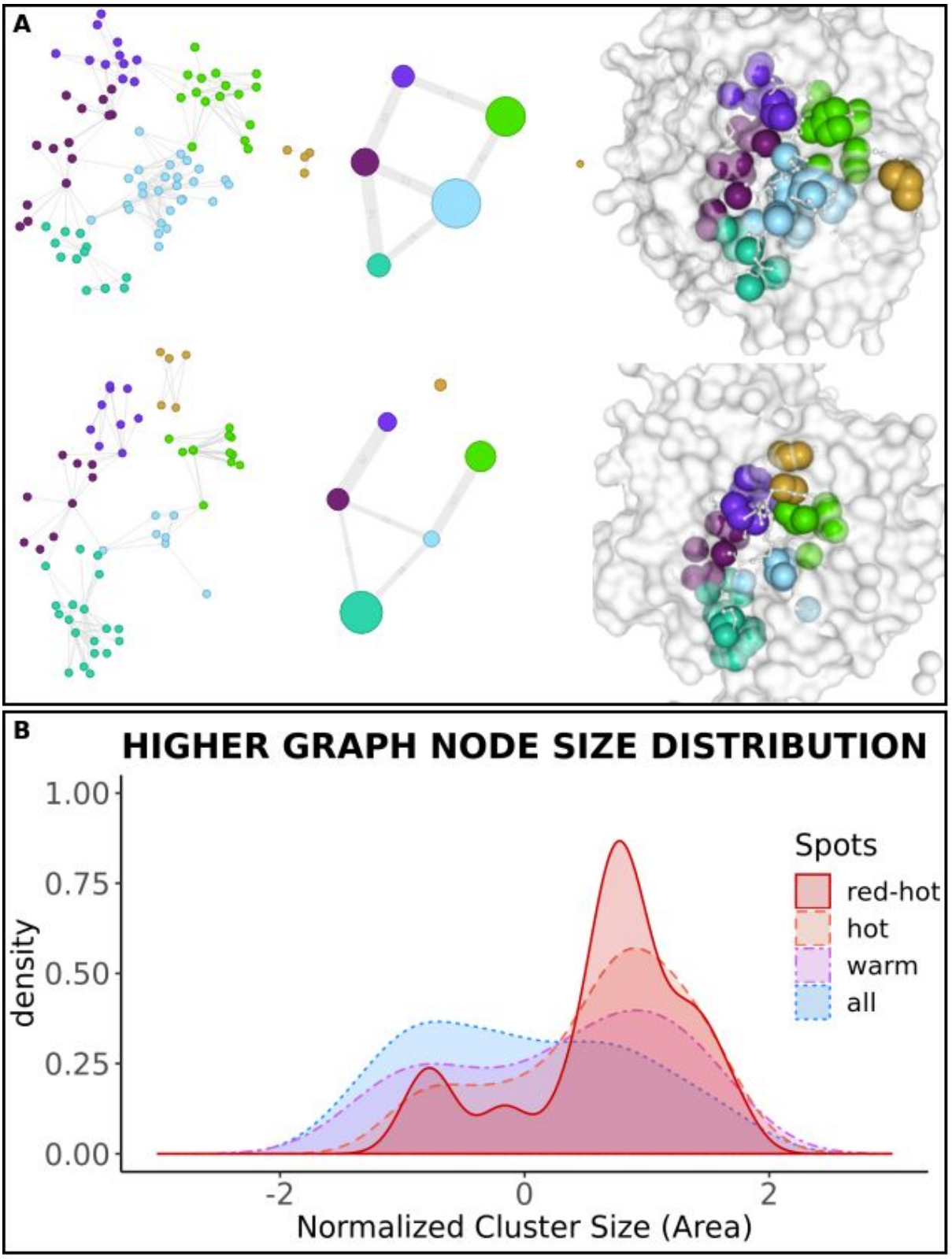

Fig. 1. (**A**) Alignment results of nonpolar interfaces from two distinct serine-peptidase in complex with also two different inhibitors (PDB ids: 1PPF and 1TEC). (**B**) Node size distribution density in function of normalized cluster sizes (z-scores of the contact areas, using a higher graph minimum quality of 0.80).