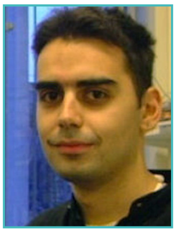
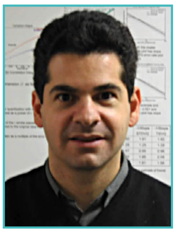
Speeding up the drug discovery process: structural similarity searches using molecular surfaces





Dimitrios Vlachakis, Georgia Tsiliki, Dimosthenis Tsagkrasoulis, Carla Sofia Carvalho, Vasileios Megalooikonomou, Sofia Kossida

Bioinformatics & Medical Informatics Laboratory, Biomedical Research Foundation of the Academy of Athens, Athens

Received 30 May 2012; **Published** 15 October 2012

Simplifying spatially complicated problems in the field of drug design, pharmacology and 3D molecular modelling is becoming very important, owing to the rapid increase in genomic and structural database sizes. The computational load is immense, and novel innovative approaches are sought, in order to perform comprehensive structural studies and 3D searches at only a fraction of the original time required (Gerld *et al.*, 2011).

Protein docking (PD) and protein-protein interactions (PPI) are two of the most rapidly emerging fields in modern structural bioinformatics. Many studies attempt to justify biological activity and function of small molecules, macromolecules or even molecular complexes using PD and PPI. For example, the majority of the information we have about the molecular processes that take place in the nucleus or the cytoplasm, and affect DNA replication, has been acquired by fast algorithms and machine-learning approaches that investigate protein-protein interactions. Molecular dynamics, genetic and epigenetic networks, systems biology, molecular biology and many other related disciplines use PD and PPI as key research tools. Many databases have been developed in this direction: e.g., the MIPS mammalian protein-protein database, the eF-site molecular surface database, the STRING database of functional protein association networks, BioGRID, VASP, PESDserv and many more (Pagel *et al.*, 2005; Kinoshita and Nakamura, 2003; Szklarczyk et al., 2011; Stark *et al.*, 2011; Chen and Honig, 2010; Das *et al.*, 2010). However, the limitation is that these approaches are modelled simulations using graph-theoretical methods, whose sensitivity and specificity is not always trustworthy. Eventually, human input and insight is required, as the application of current algorithms to all available data is impossible owing to hardware- and time- limitations.

Here, we present a novel strategy to perform similarity searches and molecular docking experiments using protein molecular surfaces. Our approach starts by calculating a series of distinct molecular surfaces for each protein, which are subsequently flattened out, thus reducing 3D information to 2D. Multiple surfaces may be combined to establish 2D Molecular Profile Fingerprints (2DMPFs) unique for each protein. 2DMPFs still retain the original 3D structural information of each protein, and may be analyzed via image-processing and pattern-recognition techniques using sliding windows and similarity-scoring functions. Finally, using fast Fourier transformation algorithms we can move from 2D image data to 1D graph lines, which are unique to each protein and can be used as fingerprints for similarity searches.

The 3D shape, size and surface information of a protein can be depicted using molecular surfaces (Nimrod et al., 2009). There are many different types of molecular surface, the commonest being electrostatic, pocket, lipophilic, b-factor and secondary-element surfaces (Binkowski and Joachimiak, 2008; Yin *et al.*, 2009; Sael *et al.*, 2008, Brylinski and Skolnick, 2010). The first task of our approach will be to calculate a set of fine-grid surfaces of each protein structure available in the RCSB Protein Data Bank (Rose et al., 2011). Then the projection of the protein surfaces from 3D to 2D representation will begin by mapping the molecular surfaces into spherical surfaces of radius proportional to the size of the protein. The resolution block (pixel) is associated with a physical size, and thus has a fixed size for all spherical surfaces. For this step, we will use the SPHARMAT package (Shen and Makedon, 2006). The second step consists of projecting the spherical

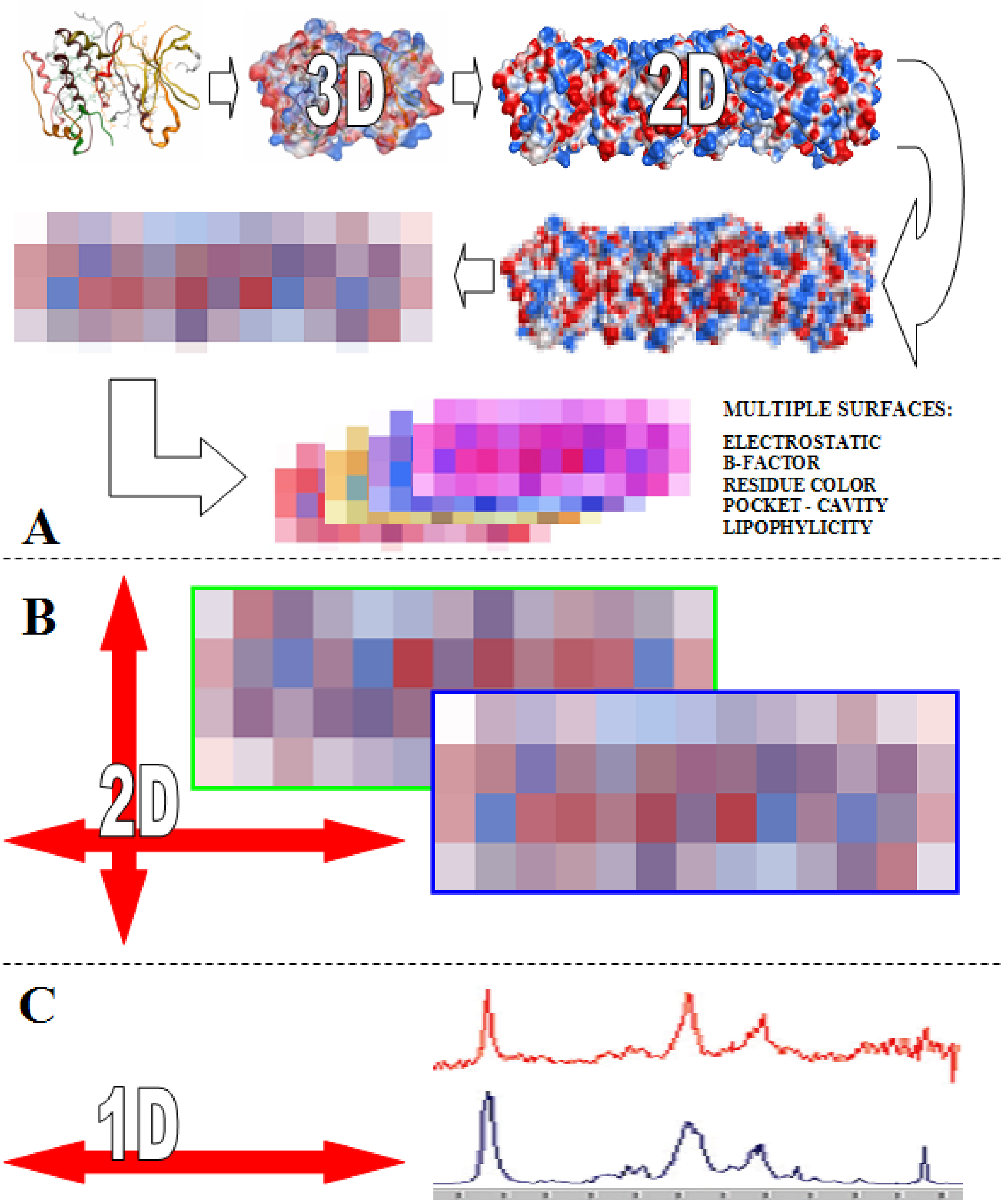


Figure 1. A schematic representation of our proposed approach. (A) The 2D protein profile fingerprint approach. The multiple surface overlapping principle, where various molecular surfaces are combined in a multiple-layer 2D image. (B) The 2D image sliding-surface pattern recognition of either matching or complementary regions. (C) Using fast Fourier transformations we can convert a 2D image to a 1D graph for even faster and more efficient pattern-recognition performance.

surfaces into flat patches along common symmetry axes of the proteins. For this step, we will use the HEALPix package (Gorski et al., 2005). The flat patches will then be the input objects for the measurement of correlations and the search of patterns among proteins. Here, multiple surfaces are being combined (Fig. 1A) and sliding 2D techniques are used for pattern recognition (Fig. 1B).

The actual scanning and filtering of the 2D data for similarity or shape/size complementary patterns will take place in the final step of the algorithm. At this stage, a correlation will be made between the results of the 2D scanning and the biological question. Multiple surfaces will have to be combined using pattern-recognition 2D sliding methodologies. The ultimate objective of our approach is to enable users to explore the computationally demanding 3D conformational space of biomolecular structures using 2D or even 1D data, which will speed up the computational process by reducing data load, without any compromise in protein information. The 1D fingerprint of our 2D images will be obtained by computing the 2-point correlation function of the Fourier transformed 2D images, which still correlate to the original 3D structure. Rather than exploring all the 3D conformational space of large protein structures when performing docking experiments, our approach will be capable of comparing the 2D image fingerprints or 1D Fourier transformed graphs (Fig. 1C) of the given structures, and in a fraction of the original time, returning results that still contain the original 3D structural information.

Multiple studies have been conducted using various correlation measures to identify patterns in 2D data (Xiong and Zhang, 2010). While working well for small datasets, the heterogeneity introduced from increased sample size inevitably reduces the sensitivity and specificity of those approaches. For this reason, we propose a model-based, pattern-recognition algorithm built under a partition-model framework, which is robust against sporadic outliers. Specifically, we assume that each 2D protein profile fingerprint can be presented by an MxN data matrix, where M is the total number of the vertical image resolution blocks and N is the total number of the horizontal ones. 2D data resolution values are categorised based on their individual value range in a scale of [-ε, ε]: ε is a positive integer empirically derived by simulated data to account for data variability. For each pair of 2D protein fingerprints, we define the difference matrix Z = {z_{ij}, i=1, ..., M and j=1, ..., N}. In this context, zij corresponds to the difference of the values in the corresponding (i,j) cell of the pair of 2D data-files, and z_{ij} = 0 if categorical values of the (i,j) cells are identical. The window could slide towards both horizontal and vertical directions, resulting in multiple similarity estimates at each pairwise comparison. The optimal window size will be estimated by minimising the Bayesian Information Criterion (BIC) of the suggested 2-way partition model (Denison *et al*., 2002). Nested partition models will be also considered. We believe that the multiple overlapping windows solution will allow us to zoom in on the 2D data in a time-inexpensive way, weight their similarities and complementarities by averaging over different neighbourhoods of the data or across data matrices, and also estimate their variability errors. Special care should be given to models' sensitivity to the categorisation scheme and estimation of optimal window size. The suggested approach will be compared with colour similarity metrics along with standard clustering techniques.

In conclusion, our approach introduces a novel technique for searching, evaluating and scoring pattern similarities between a given set of molecular surfaces. Upon calculation of a variety of diverse surface types for each protein, all 3D structural information is converted into a combined, multi-layer 2D image, which can be further simplified to 1D data via Fourier transformation. In this way, we optimise and speed up the time- and CPU-demanding 3D conformational searching, by faster more versatile 2D or 1D datasets, without compromising 3D structural information.

References

1. Binkowski TA and Joachimiak A. (2008) Protein Functional Surfaces: Global Shape Matching and Local Spatial Alignments of Ligand Binding Sites. BMC Struct Biol. 8: 45. doi: 10.1186/1472-6807-8-45
2. Brylinski M and Skolnick J. (2010) Comparison of structure- and threading-based approaches to protein functional annotation Proteins. 78, (1), 118–134. doi: 10.1002/prot.22566
3. Chen BY and Honig B (2010) VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity. PLoS Comput Biol. 6(8). doi: 10.1371/journal.pcbi.1000881

4. Das S, Krein MP, Breneman CM. (2010) PESDserv: a server for high-throughput comparison of protein binding site surfaces. Bioinformatics. 26(15), 1913–1914. doi: 10.1093/bioinformatics/btq288
5. Denison DGT , Adams NM, Holmes CC, Hand DJ. (2002) Bayesian partition modeling. Comput Stat Data An. 38, (4): 475-485. doi: 10.1016/S0167-9473(01)00073-1
6. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbit PC et al. (2011) The enzyme function initiative. Biochem. 50, 9950-9962. doi: 10.1021/bi201312u
7. Górski KM, Hivon E, Banday AJ, Wandelt BD, Hansen FK et al. (2011) HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. The Astrophysical Journal 622, 759-771. doi: 10.1086/427976
8. Kinoshita K and Nakamura H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci. 12(8), 1589–1595. doi: 10.1110/ps.0368703
9. Shen L, Makedon FS. (2006) Spherical mapping for processing of 3-D closed surfaces. Image and Vision Computing 24, (7):743-761. doi: 10.1016/j.imavis.2006.01.011
10. Nimrod G, Szilágyi A, Leslie C, Ben-Tal N. (2009) Identification of DNA-Binding Proteins Using Structural, Electrostatic and Evolutionary Features. J Mol Biol. 10, 387(4), 1040-1053. doi: 10.1016/j.jmb.2009.02.023
11. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I et al. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics. 21(6):832-834. doi: 10.1093/bioinformatics/bti115
12. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D et al. (2011). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39, D392-401. doi: 10.1093/nar/gkq1021
13. Sael L, La D, Li B, Rustamov R, Kihara D. (2008) Rapid comparison of properties on protein surface Proteins. 73, (1), 1–10. doi: 10.1002/prot.22141
14. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 39, D698-704. doi: 10.1093/nar/gkq1116
15. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39, D561-568. doi: 10.1093/nar/gkq973
16. Xiong Z and Zhang Y (2010) A critical review of image registration methods Inter J Image Data Fusion. 1, (2): 137-158. doi:10.1080/19479831003802790
17. Yin S, Proctor EA, Lugovskoy AA, Dokholyan NV. (2009) Fast screening of protein surfaces using geometric invariant fingerprints. Proc Natl Acad Sci U S A. 29, 106(39), 16622–16626. doi: 10.1073/pnas.0906146106