Research Article

3D structural analysis of proteins using electrostatic surfaces based on image segmentation

Dimitrios Vlachakis[1,2,3#], Spyridon Champeris Tsaniras[2,4#], Georgia Tsiliki[1#], Vasileios Megalooikonomou[3] and Sophia Kossida[1]

# These authors have contributed equally to this study
[1]Biomedical Research Foundation of the Academy of Athens, 11527, Athens, Greece
[2]Bionetwork ltd. 15234, Chalandri, Athens, Greece
[3]Computer Engineering and Informatics Department, School of Engineering, University of Patras, 26500 Patras, Greece
[4]Department of Physiology, Medical School, University of Patras, 26500 Patras, Greece

Received on February 3, 2014; Accepted on February 14, 2013; Published on February 28, 2014

Correspondence should be addressed to Dimitrios Vlachakis (dvlachakis@bioacademy.gr) and Sophia Kossida (skossida@bioacademy.gr)

Abstract

Herein, we present a novel strategy to analyse and characterize proteins using protein molecular electrostatic surfaces. Our approach starts by calculating a series of distinct molecular surfaces for each protein that are subsequently flattened out, thus reducing 3D information noise. RGB images are appropriately scaled by means of standard image processing techniques whilst retaining the weight information of each protein's molecular electrostatic surface. Then homogeneous areas in the protein surface are estimated based on unsupervised clustering of the 3D images, while performing similarity searches. This is a computationally fast approach, which efficiently highlights interesting structural areas among a group of proteins. Multiple protein electrostatic surfaces can be combined together and in conjunction with their processed images, they can provide the starting material for protein structural similarity and molecular docking experiments.

Introduction

It is worth noting that more than 90% of drugs tested on humans fail due to unpredicted toxicities and insufficient bioavailability properties (Kola & Landis 2004). Moreover, the mission of scientists in the post-genomic era has reached unprecedented heights that are impossible to meet using even state-of-the-art bioinformatics tools. Extra effort and funds are currently being invested to improve and speed-up the processing potential of many computer-based tools that reign in the field of structural bioinformatics. However, the underlying principle for the majority of these tools remains the same; all structural comparisons are being made mostly on a protein primary sequence identity/similarity basis. On the contrary, there are few, more advanced tools that perform structural similarities using the actual 3D information, based on the spatial coordinates of atoms within the protein structure (MOE CCG). Even though using spatial data to compare proteins is a huge leap ahead compared to the sequence-based approaches, such methodologies are slow and quite impractical to use in large-scale real-life experiments.

Exploring the 3D space of multiple enzymes that are treated as fully flexible entities requires immense processing capabilities and infrastructure. Evolutionary relationships of proteins, protein structure–function predictions and comparative modeling should all be based on searches and databases containing structural information. There are many examples of protein function annotation, where sequence based searches are insufficient (Dobson *et al.* 2004). For instance, most RNA viruses, even though they can be evolutionary linked, share very low sequence identities among homologous proteins. This is due to the fact that RNA viruses are highly mutagenic (Vlachakis 2009). Homologous proteins are more conserved in their structures than primary amino acid sequences (Illergard *et al.* 2009). Even though long studies have been carried out in areas such as structural flexible alignment and this problem has long ago been identified, it has not been yet satisfactorily addressed (Dobson *et al.* 2004, Illergard *et al.* 2009, Kolodny *et al.* 2005). The same applies to the metagenomic data, where scientists are struggling to keep up with the in-

creasing volume of biological information.

Being able to annotate a series of genes based on a sequence that can then be blasted against dedicated databases for hits in regards to their theoretical structural features, or to perform ultra-fast comparison among diverse structures of proteins, would greatly speed up the annotating bottleneck that bioinformatics currently impose on the fields of genomics and proteomics. It has been estimated that the unprocessed generated data per sequencing machine can be of the order of at least 30 Gb per day, which can scale up by a relevant factor in the case of mapped/processed data. There is a clear requirement for fast and efficient analysis of whole-genome / proteome sequencing data in the upcoming era of personalized medicine (Vlachakis et al. 2012). Due to the continuous improvements in sequencing technologies and proteomic methodologies, the current scaling of available storage capabilities and throughput analysis is limited compared to the scaling of the data generation rate. The induced lag between storage and processing potential and the corresponding requirements already poses problems for researchers and companies in the bioinformatics field. Therefore, well-defined algorithms offer a better scaling to analyze the ever increasing data (Krissinel 2012).

Why study electrostatic surfaces?

Undisputedly, the biological information contained in three dimensional structures is invaluable when studying or comparing proteins (Balatsos *et al.* 2012). Albeit, it poses a heavy burden when it comes to processing high throughput tasks (i.e. similarity searches). On these grounds, there is need for new methodologies that simplify the demanding and complex processes behind 3D protein structural comparisons (Sellis *et al.* 2009, Vlachakis *et al.* 2013a).

In an ideal scenario, all the information available within a 3D structure would be translated into a computer-friendly dataset which could be handled and processed in a much faster and more efficient manner, while at the same time the highest possible level of information detail would be retained. The whole process should be pipelined and optimized accordingly, so that it meets the current bioinformatics analysis needs and purposes (Vlachakis *et al.* 2012).

However, the vast amount of information available in a protein structure poses several barriers in how all this information can be expressed and utilized. A backbone analysis, for instance, is very useful for comparing protein structures; it does not however provide any significant insight regarding protein-protein molecular interactions via their solvent-exposed outer surfaces (Vangelatos *et al.* 2009). This type of information could help describe proteins' functions and interactions with other molecules more realistically (Yang *et al.* 2012).

Additionally, evolution can affect two proteins by altering their physicochemical properties and structural characteristics, thus promoting either a tighter or weaker molecular function and interaction pattern (Via *et al.* 2000). This is not necessarily directly related to the structural alterations of those proteins. Despite maintaining their original overall structural differences, they might share a significant resemblance in their surface regions, thus being able to catalyze similar chemical reactions (Bork *et al.* 1993). Such proteins have probably gone through divergent or convergent evolution. During convergent evolution, structurally different proteins with distinct functional similarities may develop similar electrostatic surface properties and characteristics (Sellis *et al.* 2012). Subsequently, proteins that are not homologous might share similar binding sites (Palaiomylitou *et al.* 2008). In divergent evolution, the surface characteristics of two proteins are well-preserved in order to perform the same functions even though each protein may undergo several mutations during evolution (Kauvar & Villar, 1998, Russell *et al.* 1998). Having said that, a search or comparison among different sequences or even structures might not reveal any conservation (i.e. local sites) based on mutual surface characteristics.

Previously we have used the information on protein molecular surfaces to compute a two-point correlation function in harmonic space, thus reducing the initial three-dimensional information to a one-dimensional representation of the proteins's structure (Carvalho *et al.* 2013). In this study, we attempt to reduce the three dimensional information available in 3D protein structures via image processing in order to achieve higher efficiency both in terms of data storage and computational processing power. Herein, the ultimate aim is to focus on the heavily weight areas of the protein images and report similarity scores between protein structures, and thus enable a computationally fast scan of the protein structure data before taking into account their functional details. Our working example is the comparison of two protein families. Namely, we use a set of viral helicase and polymerase proteins. In particular, we demonstrate that by means of image preprocessing and image segmentation techniques the correlation structures between the two protein families are indicative of their origin.

Methods

For the purposes of this study, we will focus on the shape, size and charges of each protein, which can be

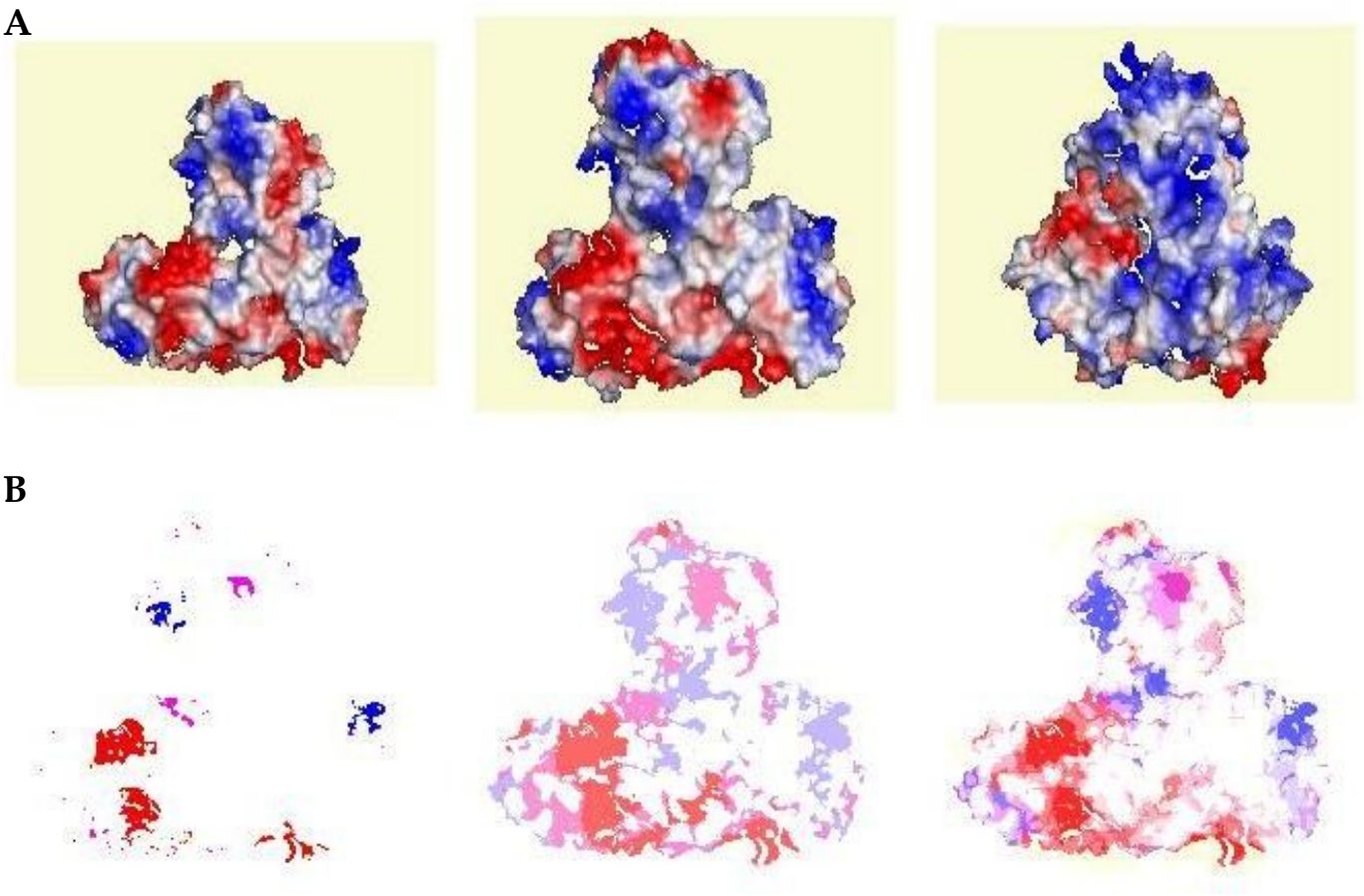

Figure 1. RGB structural plots for the proteins examined. A) RGB plots for 1A1V, 8OHM and 1NB7 proteins, respectively. B) The new protein derived by adding the RGB images of 1A1V and 8OHM proteins. The reference protein is presented after we employed three methodologies: a threshold of 50 to RGB intensities, K-means algorithm with K=3, and spectral clustering together with K-means algorithm (K=3).

displayed using vacuum electrostatic potential surfaces (Brylinski & Skolnick 2010). The first task of our approach will be to calculate a set of fine-grid electrostatic surfaces of each protein structural entry available in the RCSB PDB protein database. Then the 3D surfaces are scaled and cross normalized to be comparable between them, using techniques currently adopted in image manipulation. Each image's size will be proportional to the size of the protein, whereas grooves, channels and shapes features will be represented accordingly scaled (Figure 1). For instance, the negative and positive charges of the electrostatic surface will be represented with blue and red colour respectively, whereas white colour signifies neutral charge. The final part of the analysis is the actual scanning and filtering of the 3D data for similarity or shape/size complementarity patterns that may be of biological interest. Instead of exploring all the computationally demanding 3D conformational space of large protein structures when performing docking experiments, we are comparing the 3D raster image fingerprints (Figure 2) of the given protein structures, focusing only on the highly charged areas of the proteins considered whilst retaining the original 3D structural information (Vlachakis *et al.* 2013b).

Energy minimization of all PDB structures was done in MOE using the Amber99 (MOE CCG) forcefield as implemented into the same package. The energy minimization was set to reach the RMSD gradient of 10^{-4}, in order to ensure that any residual geometrical stereochemical strain has been removed. The model was subsequently solvated with simple point charge (SPC) water using the truncated octahedron box extending to 7 Å from the model and molecular dynamics were performed for 200 nanoseconds, at 300K/ 1 atm with a 2 femtosecond step size, using the NVT ensemble in a canonical environment (NVT stands for number of atoms, volume and temperature that remain constant throughout the molecular simulation). The results of the molecular dynamics simulation were collected into a database by MOE and can be further analyzed. Upon the energy minimization of the protein structures, their electrostatic potential surfaces were calculated by solving the nonlinear Poisson–

Figure 2. Raster images of the reference protein. Three methodologies are employed: a threshold of 50 to all intensities, K-means algorithm with K=3, and spectral clustering together with K-means algorithm (K=3). The raster grid is 25x29x3 in all cases.

Boltzmann equation using the finite difference method as implemented in the Pymol Software (DeLano 2002). The potential was calculated on grid points per side (65, 65, 65) and the grid fill by solute parameter was set to 80%. The dielectric constants of the solvent and the solute were set to 80.0 and 2.0, respectively. An ionic exclusion radius of 2.0 Å, a solvent radius of 1.4 Å and a solvent ionic strength of 0.145 M were applied. Default molecular, atomic and residue charges and atomic radii were used for this calculation.

Results and Discussion

Data preprocessing

For the purposes of this study we consider as input of the analysis the images of the 3D protein electrostatic surfaces derived from the Pymol software, which are originally preprocessed using standard image manipulation techniques. The analysis is conducted in decreased size raster RGB images, which focus on the highly charged, either blue or red, regions of the molecular surface. In Figure 1A we can observe the proteins considered from the helicase (1A1V, 8OHM) and polymerase family (1NB7). Images are scaled, averaged and normalized so that electrostatic charge colour intensities have the same variances across images and colour layers. Subsequently, RGB images are rendered into raster images (Figure 2), which allow us to further emphasize on their highly charged areas and study their 3D similarity structure. It is worth noting that the data preprocessing produces raster images of the same grid or size (namely 25x29x3). The charge intensity information lost by this procedure is negligible for the purposes of this study, because the scope of the analysis is to solely focus on the heavily charged areas (colored in blue and red) of the protein structure which serve as a protein signature and consequently as a prior criterion for similarity searches.

Unsupervised Image Segmentation

In this study we describe our initial results from 3D protein structure image segmentation by means of spectral clustering (Ng *et al.* 2001). The primary goal of the image segmentation is to estimate clusters of data points where within each cluster data are highly correlated and uncorrelated with the data in the remaining clusters. Data could then be summarized by specifying the number, size and properties of the estimated image clusters. There are many ways to partition an image, such as adaptive thresholding algorithms, local intensity gradient methods and hierarchical clustering (Theodoridis & Kountroumbas 2003, Freixenet *et al.* 2002). Here we consider spectral clustering, which is an unsupervised classification methodology often used in image analysis (Liu *et al.* 2010, Tung *et al.* 2010). By employing spectral clustering to preprocessed RGB images, we demonstrate that we are able to effectively estimate their colouring homogeneity and, thus, to classify them in the two protein families considered. The novelty in this case lies in analyzing the 2D electrostatic representation of the projected 3D protein structural surfaces.

Spectral clustering makes use of the top eigenvectors of the similarity matrix calculated among the data points to further reduce dimensionality (Ng *et al.* 2001). The flexibility behind spectral clustering, as with other clustering techniques, is that the similarity matrix between any two data points, or a neighborhood of data points, could be defined in many ways depending on the data analyzed and the similarity measure used; an example is the Euclidean distance and the kernel function of the Euclidean distance. Furthermore, by

Proteins/ Methods	Preprocessed RGB images	Threshold	K-means	Spectral Clustering
1A1V	0.3566337	0.7629626	0.8411391	0.8500275
8OHM	0.3598572	0.8081752	0.8459892	0.83574
1NB7	0.3586333	0.1490731	0.4905345	0.2750232
HCV_1HEI	0.3587256	0.7293267	0.8153672	0.8307631

Table 1. Pearson correlation coefficients between the new synthesized RGB image (1A1V + 8OHM) and proteins from helicase (1A1V, 8OHM, HCV_1HEI) and polymerase family (1NB7).

relying on the similarity between pair-wise or any other combination of data, we take into account the underlying interactions in the data. For this particular application the similarity statistic used is a Gaussian kernel of the difference between raster data points (Burt & Anderson 1983) with variance equal to 0.025, which has been empirically derived by similarly manipulated protein images. The weighted Laplacian similarity matrix is considered. Statistical analysis has been conducted using R 3.0.1 (R Core Team 2013).

Spectral clustering is combined with the K-means algorithm (K=3), and the image segmentation results are compared to the segmentation produced when we impose a threshold of 50 to all RGB intensity values, as well as employing the K-means algorithm alone. The K-means algorithm is selected as a baseline algorithm widely applied to partitioning problems. In Figure 2, we show an example of combining two segmented proteins from the same family, 1A1V and 8OHM, where we employ thresholding, K-means and spectral clustering with K-means, respectively. The three methodologies are applied separately to both of the two proteins before their two images are added. By adding the two images we only keep the commonly charged areas of the proteins. In that way, we form a reference protein, which is indicative to the structural characteristics of the protein family.

Figure 2 shows the reference protein's raster image with a 25x29x3 raster grid, generated from the segmented image shown in Figure 1B, where the red and blue areas correspond to highly charged areas. In Table 1, we present the Pearson correlation coefficients between the reference protein and the three proteins considered here. More specifically, the first column shows the correlation coefficients between the preprocessed images as shown in Figure 1A. We can observe that the three proteins are not significantly correlated with the reference protein, which is now produced by adding the two proteins as in Figure 1A. This is irrespective to the protein family. The remaining columns of Table 1 show the similarity coefficients for the raster images produced under the three methodologies mentioned above, namely thresholding, K-means and spectral clustering. We can observe that for all three methodologies, the reference protein is highly correlated to the first two proteins and uncorrelated with the 1NB7 protein; however spectral clustering seems to better distinguish the two protein families. Finally, results are reported for the HCV_1HEI protein (last row of Table 1). The HCV_1HEI protein belongs to the helicase family and is expected to be highly correlated to the reference protein. We can observe a similar correlation pattern to the 1A1V and 8OHM proteins, which is a promising result for the generalization of the methodology presented here, as these two proteins also belong to the helicase family, while 1NB7 is a polymerase protein.

Conclusions

We have implemented a novel workflow methodology for the segmentation of the protein structure images focusing on the shape, size and electrostatic charge of the protein. By means of unsupervised image segmentation we were able to distinguish the highly charged areas in the protein's image, and form a reference protein. Our goal was to search for structural molecular similarities between electrostatic surfaces in the protein domain. In this paper we have concentrated on establishing the advantages in producing a reference protein that describes the common structural characteristics of a protein family, which will then serve as a baseline for the classification of other proteins. The preliminary results presented here show that there is scope for extending the suggested methodology to a wider pool of protein data. We believe that our methodology will serve as a first search criterion for protein similarities, as the dimensionality of the data is considerably decreased allowing for time-inexpensive searches.

Acknowledgements

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Pro-

gram "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund. This work was also funded by the BIOEXPLORE research project. BIOEXPLORE research project falls under the Operational Programme "Education and Lifelong Learning" and is co-financed by the European Social Fund (ESF) and National Resources.

References

Balatsos N, Vlachakis D, Chatzigeorgiou V, Manta S, Komiotis D, Vlassi M & Stathopoulos C 2012 Kinetic and in silico analysis of the slow-binding inhibition of human poly (A)-specific ribonuclease (PARN) by novel nucleoside analogues. *Biochimie* **94** 214-221

Bork P, Sander C & Valencia A 1993 Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci.* **2** 31-40

Brylinski M & Skolnick J 2010 Q-Dock(LHM): Low-resolution refinement for ligand comparative modeling. *J Comput Chem* **31** 1093-1105

Burt P & Adelson E 1983 The Laplacian pyramid as a compact image code. *IEEE T Commun* 31 **4** 532-540

Carvalho CS, Vlachakis D, Tsiliki G, Megalooikonomou V & Kossida S 2013 Protein signatures using electrostatic molecular surfaces in harmonic space. *PeerJ* **1** e185

Dobson PD, Cai YD, Stapley BJ & Doig AJ 2004 Prediction of protein function in the absence of signicant sequence similarity. *Curr Med Chem* **11** 2135-2142

DeLano WL 2002 The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org.

Freixenet J, Muñoz X, Raba D, Martí J & Cufí X 2002 Yet another survey on image segmentation: Region and boundary information integration. *Proceedings of the European Conference on Computer Vision* **3** 408-422

Illergard K, Ardell DH & Elofsson A 2009 Structure is three to ten times more conserved than sequence a study of structural response in protein cores. *Proteins* **77** 499-508

Kauvar LM & Villar HO 1998 Deciphering cryptic similarities in protein binding sites. *Curr Opin Biotechnol* **9** 390-394

Kola & Landis 2004 Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* **3** 711-715

Kolodny R, Koehl P & Levitt M 2005 Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol* **346** 1173-1188

Krissinel E 2012 Enhanced fold recognition using efficient short fragment clustering. *J Mol Biochem* **1** 76-85

Liu HQ, Jiao LC & Zhao F 2010 Non-local spatial spectral clustering for image analysis. *Neurocomputing* **74** 461-471

MOE CCG, 1010 Sherbrooke St. West, Suite 910, Montreal, Canada, H3A 2R.

Ng AY, Jordan MI & Weiss Y 2001 On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* 849-856

Palaiomylitou M, Tartas A, Vlachakis D, Tzamarias D & Vlassi M 2008 Investigating the structural stability of the Tup1-interaction domain of Ssn6: Evidence for a conformational change on the complex. *Proteins* **70** 72-82

R Core Team 2013 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Russell RB, Sasieni PD & Sternberg MJE. 1998 Supersites within superfolds. Binding site similarity in the absence of homology. *J Mol Biol* **282** 903-918

Sellis D, Drosou V, Vlachakis D, Voukkalis N, Giannakouros T & Vlassi M 2012 Phosphorylation of the arginine/serine repeats of lamin B receptor by SRPK1-Insights from molecular dynamics simulations. *Biochim Biophys Acta* **1820** 44-55

Sellis D, Vlachakis D & Vlassi M 2009 Gromita: a fully integrated graphical user interface to gromacs 4. *Bioinform Biol Insights* **3** 99-102

Theodoridis S & Koutroumbas K 2003 Pattern Recognition. Edn 5, New York: Academic Press

Tung F, Wong A & Clausi DA 2010 Enabling scalable spectral clustering for image segmentation. *J Pattern Recogn* **43** 4069-4076

Vangelatos I, Vlachakis D, Sophianopoulou V & Diallinas G 2009 Modelling and mutational evidence identify the substrate binding site and functional elements in APC amino acid transporters. *Mol Membrane Biol* **26** 356-370

Via A, Ferre F, Brannetti B & Helmer-Citterich M 2000 Protein surface similarities: a survey of methods to describe and compare protein surfaces. *Cell Mol Life Sci.* **57** 1970-1977

Vlachakis D 2009 Theoretical study of the Usutu virus helicase 3D structure, by means of computer-aided homology modelling. *Theor Biol Med Model* **25** 6 9

Vlachakis D, Pavlopoulou A, Tsiliki G, Komiotis D, Stathopoulos C, Balatsos NA & Kossida S 2012 An integrated in silico approach to design specific inhibitors targeting human poly(a)-specific ribonuclease. *PLoS One* **7** e51113

Vlachakis D, Tsagrasoulis D, Megalooikonomou V & Kossida S 2013a Introducing Drugster: a comprehensive and fully integrated drug design, lead and structure optimization toolkit. *Bioinformatics* **29** 126-128

Vlachakis D, Champeris Tsaniras S & Kossida S 2013b Insights into the structure and 3D spatial arrangement of the b-ketoacyl carrier protein synthases. *J Mol Biochem* **2** 150-158

Vlachakis D, Tsiliki G, Tsagkrasoulis D, Carvalho CS, Megalooikonomou V & Kossida S 2012 Speeding up the drug discovery process: structural similarity searches using molecular surfaces. *EMBnet J* **18** 6-9

Yang H, Qureshi R & Sacan A 2012 Protein surface representation and analysis by dimension reduction. *Proteome Sci* **10** S1