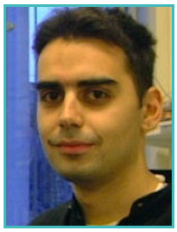
# The future of structural bioinformatics in the post-genomics era; a novel framework to facilitate modern drug design




**Dimitrios Vlachakis, Dimosthenis Tsagkrasoulis, Georgia Tsiliki, Sophia Kossida**

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

**Received** 2 May 2012; **Published** 15 October 2012

## Concept description

A major drawback of *in silico* protein science nowadays is that protein structural comparisons are based on sequence searches. Evolutionary relationships of proteins, protein structure–function predictions and comparative modelling would all benefit from greater use of structural information. There are many examples of protein function annotation where sequence-based searches are insufficient (Dobson *et al.*, 2004). Most RNA viruses, even though they can be evolutionarily linked, share very low sequence identities among their homologous proteins, as they are highly mutagenic. Even though the structures of such are more conserved than their sequences (Illergard *et al.*, 2009), and studies have been carried out in areas such as flexible structural alignment, this fact has nevertheless not yet been satisfactorily utilised (Kolodny *et al.*, 2005; Berbalk *et al.*, 2009; Mayr *et al.*, 2007).

A novel approach that exploits the immense size of genomic databases and links them to structure is presented in this study. Both major types of databases are involved in our methodology: the RCSB-PDB, a database of known biological structures, with information obtained mostly by X-ray crystallography and NMR studies (Rose *et al.*, 2011); and enormous genomic databases, such as the NCBI GenBank and Whole Genome Shotgun (WGS) databases, which contain sequence information from many species (including human) acquired by various large- and small-scale sequencing approaches (Benson *et al.*, 2012; Johnson *et al.*, 2008). At the last count, the PDB contained a total of 77,878 structures, whereas GenBank contains 126,551,501,141 bases in 135,440,924 sequence records, plus another 191,401,393,188 bases in 62,715,288 sequence records in the WGS division.

In our method, PDB structures will need no preliminary analysis, while on the other hand, the DNA sequence data-sets, bigger by several orders of magnitude, will have to undergo special filtering – this will include ruling out low complexity regions and focusing on exonic sequence space, a task that will contribute significant noise-reduction to the initial data. Notably, both major databases involved in this use case have been growing exponentially in size over the last few years (Rose *et al.*, 2011; Benson *et al.*, 2012).

The new methodology will provide the tools required to perform protein similarity searches based on structural rather than sequence information. The input query sequence can either be of known or unknown structure (Figure 1). In each case, the primary amino acid sequence will need to be converted to the amino acid Structural Features Sequence (SFS) format. The SFS format is a novel residue-annotation method based on the structural conformation of each amino acid in the query sequence. For instance, residues forming an $\alpha$-helix will be replaced with an "H", a $\beta$-sheet with an "S", a coil with "C", until all query amino acids have been designated with an SFS value. If the input sequence is of unknown structure, it will be subjected to secondary structure prediction algorithms, and the SFS format will be deduced. The same SFS formatting principle and secondary structure prediction algorithm must be applied to both NCBI databases, which can either be performed on the fly or by the one-off conversion of all known information into a new databank, which will need to be updated regularly. As all entries in the PDB contain secondary structural information, the conversion to SFS for-

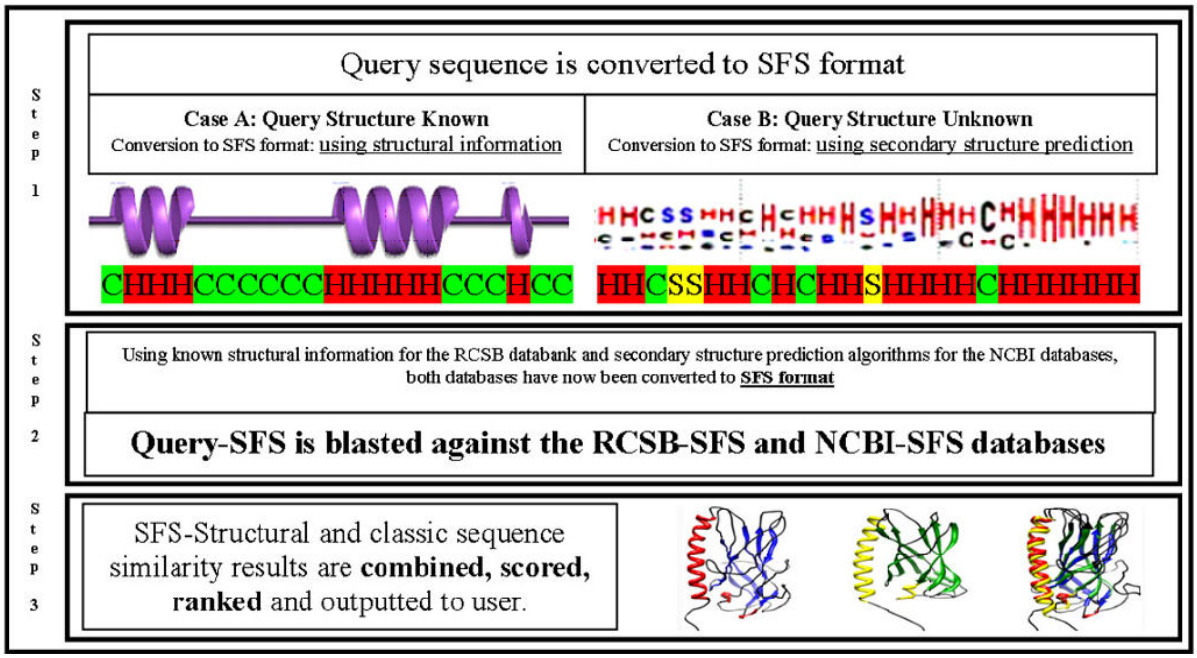



**Figure 1.** Diagram of the proposed approach.

mat can be performed without any predictions. Our proposed algorithm will be broken down into three parts: in part 1, the query sequence will be converted to SFS format; in part 2, the query-SFS will be structurally aligned against all structures in the PDB-SFS-formatted database and all sequences in the NCBI-SFS-formatted databases; finally, in step 3, structural similarity results will be combined with classic sequence BLAST results and output to the user.

Because our data are, by default, incomplete in the case of genomic sequences that lack structural information, we plan to develop and apply a fast and efficient secondary structure prediction algorithm. However, even upon application of the algorithm, it is still possible to obtain "noisy" data if the prediction score does not clearly indicate structural features. There are two different approaches to deal with this issue. The first is to use multiple secondary structure prediction algorithms, some of which are already established. By applying a variety of different algorithms and approaches on the same sequence string, we will achieve a 'consensus prediction' that will be statistically more reliable. Secondly, we plan to develop a clever algorithm that we will train to recognise and annotate the origin and function of each unknown DNA sequence string using Artificial Intelligence (AI) and machine-learning techniques. Then, by 'homology and comparative approaches', we will be able to 'predict and expect' various structural elements in a given sequence and, accordingly, adjust the weight ratios used by the secondary prediction algorithm. For example, if we obtain noisy/unclear data from the exonic product of a DNA sequence that has been found to contain conserved features of a certain family of transcription factors with $\alpha$-helical repeats, then the algorithm will 'expect' that sequence to have similar $\alpha$-helical conformation. It is important to clarify that the 'consensus prediction' and the 'homology and comparative approaches' will only be applied when noisy data appear, saving CPU calculation effort when the data are clean.

## Outlook

The real world problem addressed by our new methodology is highly relevant to the general field of biomedicine. Providing a concise and efficient framework for detecting protein structural similarity is bound to be very valuable for experimental drug design. Almost 90% of drugs tested on humans fail owing to unpredicted toxicities. Supplying the bio-pharmaceutical industry with a compendium of easily searchable and retrievable structures against which any substance of interest may be compared in a straight-forward

manner, will enable the filtering out of a significant amount of probable side-effects. This would imply increasing the expected effectiveness of the proposed drug with a simultaneous significant decrease in cost. The pharmaceutical industry would benefit enormously in fields such as drug design and development, by being able to search for similar structural features and active sites for a given drug or inhibitor.

## References


1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. *Nucleic Acids* Res **40**, D48-53. doi: 10.1093/nar/gkr1202
2. Berbalk C, Schwaiger CS, Lackner P (2009) Accuracy analysis of multiple structure alignments. *Protein Sci* **18**, 2027–2035. doi: 10.1002/pro.213
3. Dobson PD, Cai YD, Stapley BJ, Doig AJ (2004) Prediction of protein function in the absence of significant sequence similarity. *Curr Med Chem* **11**, 2135–2142.
4. 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. doi: 10.1002/prot.22458
5. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S et al. (2008) NCBI BLAST: a better web interface. *Nucl. Acids Res* **36**, W5-W9. doi: 10.1093/nar/gkn201
6. Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol* **346**, 1173–1188. doi: 10.1016/j.jmb.2004.12.032
7. Mayr G, Domingues FS, Lackner P (2007) Comparative analysis of protein structure alignments. *BMC Struct Biol* 7: **50**. doi: 10.1186/1472-6807-7-50
8. 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