

# TOPOFIT-DB and T-Server: A Database of structural alignments and a server for one-to-all protein structure comparisons based on the TOPOFIT method

Chesley M. Leslin<sup>1</sup>, Alex Abyzov<sup>1</sup>, and Valentin A. Ilyin<sup>1</sup>

**Keywords:** protein structure, structure alignment, structural similarity, topological invariant, common structural core, structural alignment database, structural neighbor, structural representative

Studies on protein structural alignments address several questions: functional characterization and annotation of proteins, evolutionary relationship between species, identification of rigid and flexible parts, identification of invariant and variable regions, domain identification, protein engineering, modeling, molecular dynamics, etc. To accomplish a specific goal, one may need to perform only several structural pairwise comparisons which can be done in one of many applications, for example, in Friend [1]. Nevertheless, there are many research areas, which require comparison of one particular structure with many or all known protein structures. A significant number of alignment methods have been developed, and a number of popular and very useful databases of structural neighbors are available to the public. But the problem of structural comparison remains unsolved; moreover, the results from different popular databases overlap at only 40% [2], while the amount of new structures is growing. Advances in the Protein Structure Initiative (PSI), aimed at significantly reducing the costs and time it takes to determine a three-dimensional protein structure, promise an abundance of protein structures available to researchers for analysis. Therefore, there are needs for a diversity of comprehensive databases of structural alignments, clustered into appropriate representatives, linked with the appropriate tools for the visualization and analysis of the copious amounts of data.

The common strategy to align protein structures is to find the best solution by balancing between maximizing the size of the alignment, i.e. the number of aligned residues ( $N_e$ ) and minimizing the average distance between them, so called RMSD (root mean square deviations). Recently we have discovered that it is possible to objectively identify a clear border between the structural invariant and variable parts in protein structure by comparative analysis of the topologies of the Delaunay tessellation patterns, using the topomax point defined by the TOPOFIT method [3] innovative aspect of this approach is to objectively find the common structural core identified by the topomax point without subjective balancing between larger  $N_e$  and smaller RMSD. Thus, TOPOFIT identifies not only the invariant structural core between proteins but also clearly shows the variable parts. The method provides new insight in protein structure analysis with the possibility to perform a more detailed analysis and classification of those variable parts.

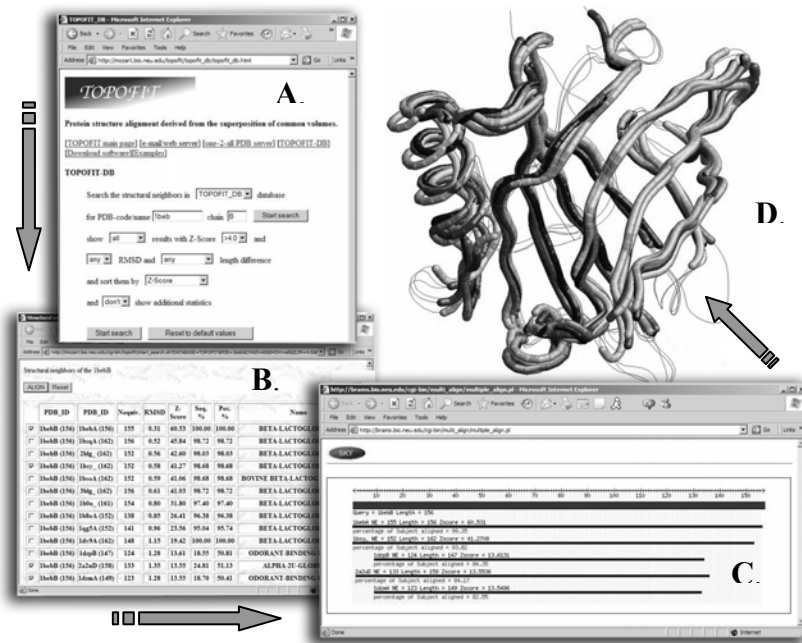
TOPOFIT-database (T-DB) is an ongoing research project, which is available to the public for search, and presently has more than 3 million alignments pre-calculated. The database scheme is similar to both CE [4] and DALI [5] structure alignment databases. Fig 1A. The 1<sup>st</sup> search page is for submission of a query PDB structure code and allows users to collect related structures by a number of cutoff parameter; alignment length, Z-score, and RMSD, sort the results by any of the above parameters, and select whether to show all statistics or a subset found in the database. Fig 1B. Displays the structurally related proteins identified by TOPOFIT, with all data shown in a table, from here users can download a tab delimited version of the data for use in Excel, or select which proteins they would like to visualize in a multiple structure alignment, with all selected

---

<sup>1</sup> Department of Biology, Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA.  
E-mail: [ilyin@mozart.bio.neu.edu](mailto:ilyin@mozart.bio.neu.edu)

proteins aligned to the query protein. Fig 1C. displays the graphical representation of the alignment with data about each alignment (Ne, Z-score, Length, and % of subject aligned to query). Fig 1D. shows the structure window in Friend, displaying the multiple structural alignment.

TOPOFIT-Server (T-Server) provides a relatively quick and straightforward way to locate structurally similar proteins in a one-to-all manner, comparing the user submitted protein structure against the entire PDB, which is updated weekly. T-server is publicly available and has an average search time of ~3hours, depending on the size of the query protein and location in the job queue. Processes are run on a fast 20 node dual processor cluster with notifications of submission and completion through email. Results are easily viewed by our sequence and structure viewer Friend in a one-to-all multiple structural alignment based upon how the subjects aligned to the submitted protein. All results are also available for download in a number of formats. T-server and T-DB are publicly available from <http://mozart.bio.neu.edu/topofit>.



**Figure 1.** View of T-DB. 1). Query page 2). Results page 3). Alignment Page 4). Multiple alignment shown in Friend structure sequence viewer (structure window displayed only).

## References

- [1] Abyzov, A, Leslin, C., and Ilyin, V. A. 2003. Friend, An Integrated Analytical Front-end Application for Bioinformatics. *Computational and Systems Biology at MIT (CSBI 2003)*.
- [2] Shindyalov IN, Bourne PE. 2000. An alternative view of protein fold space. *Proteins*. 38:247-260.
- [3] Ilyin V.A., Abyzov A, Leslin C. 2004. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Science*. 13:1865-1874.
- [4] Shindyalov IN, Bourne PE. 2001. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Research*. 29:228-229
- [5] Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*. 233:123-138