**DATE: 18-03-22**

**WEBLEM 7**

**Introduction to Structural Blast – VAST and DALI**

The protein structures that populate the PDB have provided crucial insights at the atomic level as to the molecular mechanisms that underlie protein function. Indeed, structural studies have had, and continue to have, a significant and sometimes revolutionary impact in all areas of biology. However, structural biology has tended to focus on single proteins or biological systems and, despite significant advances in the general area of structural bioinformatics, the horizontal integration of the vast quantity of structural information available in the PDB has had little or no impact in the larger biological community. This is in contrast to protein sequence information, which is more routinely, automatically and broadly used. Given that structure is more conserved than sequence, structural similarity has the potential to yield a great deal of functional information that sequence relationships cannot provide and to identify relationships between many more pairs of proteins. In this article we argue that the exploitation of statistical and machine learning techniques combined with the vast amount of high-throughput experimental data constantly being generated enable a significant expansion in the scale and diversity of application of structural information to biological problems.

The ultimate potential impact of both global and local structural relationships in inferring function is highlighted by the observation that, given a suitably “loose” definition of structural similarity, the repertoire of structures currently in the structural databases is nearly complete at the domain level. Thus, it can be expected that most newly solved protein structures will have both near and remote structural neighbors which can provide clues as to their function. Programs such as BLAST use local sequence relationships to quickly scan sequence databases. Since structure-based scans of protein structural databases can be carried out very quickly with current technology (typically minutes for a database of tens of thousands of structures), a similar strategy can be used for structural relationships as well, essentially defining a “structural BLAST”

Comparative analyses of protein sequences and structures play a fundamental role in understanding proteins and their functions. Assuming an evolutionary continuity of structure and function, describing the structural similarity relationships between protein structures allows scientists to infer the functions of newly discovered proteins. The most widespread purpose of structural alignment has been to identify homologous residues (encoded by the same codon in the genome of a common ancestor). Mutations manifest in plastic deformations, shifts and rotations of the secondary structure elements (SSEs). A wide spectrum of structural alignment methods exist, which differ in their treatment of structural variations, scoring functions and optimization algorithms.

There are aware of half a dozen web servers that provide structure comparisons against the current weekly updated Protein Data Bank (PDB). Each server is unique because they employ different structure comparison methods.

**VAST:**

The VAST search database and database of precomputed structure alignments have been maintained as complete and redundant collections since their launch, with automated updates occurring on a weekly basis. This was made possible by implementing a fast heuristic that uses a model for the statistical significance of initial alignments of secondary structure vectors (which can be computed quickly), so that the database searches can avoid costly alignment refinements for the large majority of insignificant and uninteresting similarities. The drawbacks are that a heuristic will miss some potentially interesting similarities. The VAST algorithm will not, for example, report similarities between structures deemed to have secondary structure elements. Searches for structural similarity can and should be complemented with searches for sequence similarity, as flexibility of molecular structure and limitations of the structure comparison method may preclude the detection of matches between structures of homologous polypeptides. In general, though, structure comparison methods will pick up many subtle similarities that evade detection by sequence comparison strategies, and there is no natural cutoff point for a ranked list of similar structures, unlike in the sequence comparison scenario, where matches to non-homologous gene products are considered accidental and uninformative, for the most part.

Results computed by the VAST algorithm have been compared against other approaches a number of times. Although there are subtle differences in retrieval sensitivity and alignment accuracy, it appears fair to state that the large majority of extensive structural similarities, which are indicative of common evolutionary descent and could be used to infer functional similarities, are reported by VAST (and by most if not all of the alternative approaches to detect common substructures).

As structure similarity search strategies have been developed to also detect distant relationships that might not be evident from sequence analysis, most if not all of the current approaches have been implemented so that they use a single protein molecule or rather a single domain as the unit of comparison. This has been true for VAST, in particular. However, the Protein Data Bank is continuing to accumulate structures of larger macromolecular complexes and has started to provide data on what constitutes functionally or biologically relevant macromolecular complexes or biological assemblies. Such assemblies range from simple homo-oligomers to intricate arrangements of many different components, revealing details on specific molecular interactions and on how these might constrain sequence variation. A small number of approaches have been published in the past few years that examine structural similarity of macromolecular complexes. Here we present a simple strategy that builds on the existing database of pairwise structure alignments computed by VAST and supports the first (to our knowledge) comprehensive and regularly updated collection of macromolecular complex similarities.

**VAST+ as an extension to existing protein structure comparison**

As information characterizing biological assemblies in macromolecular structure data has become available, it seemed that the biological assembly would be a convenient and informative unit of comparison between individual entries in the structure database. If the goal is to list structures most similar to any particular query, one would have to consider that the query itself may contain a macromolecular complex with a given stoichiometry, and that matching complexes with matching stoichiometry might be more informative ‘structure neighbors’ than, for example, the structures that happen to contain molecules with the strongest local similarity to the query, irrespective of the context.

VAST+ builds on the existing VAST database to generate such a report of structure neighbors. Its goal is to find the largest set of pairs of matching macromolecules between two biological assemblies and to characterize that match and compute instructions for a global superimposition that can be used to visualize the structural similarity. For each pair of structures in MMDB, VAST+ examines pre-computed structure alignments stored in the VAST database that were computed for the full-length protein molecule components of the default biological assemblies. If such pairwise alignments are found, the alignments between individual protein components of the biological assemblies are compared with each other for compatibility, and compatible/matching alignments are clustered into sets of alignments that together constitute a biological assembly match. Pairwise alignments are compatible (i) if they do not share the same macromolecules, i.e. a protein molecule from one assembly cannot be aligned to two molecules from the other assembly at the same time and (ii) if they generate similar instructions (spatial transformation matrices) for the superpositions of coordinate sets. A simple distance metric can be used to compare transformation matrices and it lends itself to cluster alignment sets efficiently.

Each set of compatible pairwise alignments can be characterized by (i) the number of pairwise matches, i.e. the total number of pairs of protein molecules from the query and subject biological assemblies, that are simultaneously aligned with each other; (ii) the RMSD of the superposition obtained from considering all alignments in the set; (iii) the total length of all pairwise alignments, i.e. the total number of amino acids that are aligned in 3D space; and (iv) percentage of identical residues in the alignments. For each pairwise comparison of two biological assemblies, only the match with the highest number of aligned molecules and the highest number of aligned residues is recorded and reported.

Currently, 53% of polypeptide-containing structures in MMDB have >1 polypeptide chain. The histogram plotted in Figure 1 breaks down the numbers by oligomer size and indicates that large fractions of the oligomeric assemblies have, in general, structure neighbors that match the entire assemblies. It should be noted that the fractions might be somewhat exaggerated, as exact duplicates of a structure would be counted as biological assembly matches, and no attempt was made to remove redundant structures or classify biological assembly matches as informative versus uninformative.

**DALI:**

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

User can perform three types of database searches:

* Heuristic PDB search - compares one query structure against those in the Protein Data Bank
* Exhaustive PDB25 search - compares one query structure against a representative subset of the Protein Data Bank
* Hierarchical AF-DB search - compares one query structure against a species subset of the AlphaFold Database

There are two types of structure comparisons of user selected structures:

* Pairwise structure comparison - compares one query structure against those specified by the user
* All against all structure comparison - returns a structural similarity dendrogram for a set of structures specified by the user

**DESCRIPTION OF THE SERVER**

**Inputs**

The input to the server is one or two protein structures in PDB format. The query structure can be specified as a PDB identifier plus chain identifier, or a PDB file uploaded by the user. There are three cross-linked query forms for the Dali server, Dali Database and pairwise comparison, respectively. For example, the entry point to the Dali server is <http://ekhidna.biocenter.helsinki.fi/dali_server>.

All backbone atoms (N, CA, C, O) are required and the minimum chain length is 30 amino acids. Backbone atoms may be reconstructed from a CA trace using the MaxSprout server at.

External links to the Dali database should use , where 1nnn represents a PDB identifier and chainid is optional. Meta-servers may link to, which directly returns the match list and alignment data as plain text.

**Processing**

Queries to the Dali Database and pairwise comparison are processed interactively; the result is usually returned within a minute. The Dali server processes up to eight PDB searches in parallel, others are queued. Most PDB- search queries are processed in less than an hour. Results are stored on the server for two weeks. The results of identical queries are retrieved instantly from cache.

The Dali server and Dali database return only the best match of the query to each PDB structure. The pairwise comparison returns also suboptimal matches. The pairwise comparison is based on a systematic branch-and- bound search that returns non-overlapping solutions in decreasing order of alignment score. Suboptimal matches can be of interest in cases of internal symmetries or repeated domains.

Dali Database is updated twice a year and contains precomputed structural alignments of PDB90 against the full PDB. The query structure is mapped to the closest representative in PDB90 and the structure comparison scores are recomputed using the transitive alignment via the representative.

The Dali server aims to retrieve a list of 500 structural neighbors of the query structure with the highest Z- scores. Most query structures have strong similarity to a structure already in the PDB. We use fast filters to identify a shortlist of about 100 promising candidates. If these produce strong matches, the search proceeds by walking. Otherwise, the query structure is compared with PDB90 in one versus all fashion, followed by a walk to collect matches to redundant PDB structures (which are over 90% sequence identical to PDB90 representatives).

Walking selects targets for structural comparison from the neighbours of neighbours found so far. The second shell of neighbors is known because all structures in the PDB are stored in a precomputed network of similarities. The pairwise alignments (Q,P) and (P,R) induce a transitive alignment (Q,R), which is used as the starting point of refinement rather than optimizing the alignment from scratch. There are many possible choices of intermediate structure P en route from Q to R. We select the ‘high road’, in other words, the minimum of the Z-scores Z(Q,P) and Z(P,R) should be as high as possible. The ‘high road’ may change as more structures are added to the first neighbour shell. To avoid redundant comparisons, we only test induced alignments which are longer than previously obtained ones. When the alignment (Q,R) has been refined, R is added to the first neighbour shell. The walk ends when either there are no new neighbours in the second shell, a specified number of hits (1000) have been reported, or a maximum number of comparisons (1000) have been performed.

**Outputs**

The Dali server, Dali Database and pairwise comparison use a common output format and share interactive analysis tools.

The result consists of (i) a list of structural neighbours, ranked by Z-score, and (ii) the alignment data. The results are presented as plain text for downloading by downstream application, and as hypertext for interactive analysis. The default results page reports the top 500 matches to all chains in the PDB. A subset of matches to PDB90, filtered at 90% sequence identity, is provided for convenience.

Selected subsets of matches can be visualized (i) as multiple sequence alignments, or (ii) in multiple 3D superimposition. While sophisticated tools with integrated sequence alignment and structure superimposition views are available, we have chosen Jmol, an open source Java viewer for molecular graphics, because it was most easily accessible to the casual user. Each neighbour is aligned (superimposed) against the query structure in a star-like tree topology. Active sites can be recognized by clusters of conserved residues and ligands. Sequence and structure conservation are calculated within the selected subset of matches.

VAST and DALI are thus very useful structure similarity BLAST tool. VAST provides user with similar structures to their query along with its molecular components and chemicals and non-standard biopolymers, aligned sequences and 3D structure superimposition information which includes information regarding H- bonds, interactions, buried surface area, 2D interaction network and much more. DALI provides user with similar structures to their query along with its pairwise alignment, coordinates information, 3Dsuperimposition results. Describing the structural similarity relationships between protein structures allows scientists to infer the functions of newly discovered proteins.

**REFERENCES:**

1. Dey, Fabian; Cliff Zhang, Qiangfeng; Petrey, Donald; Honig, Barry (2013). Toward a “Structural BLAST”: Using structural relationships to infer function. Protein Science, 22(4), 359– 366. doi:10.1002/pro.2225
2. Madej, T.; Lanczycki, C. J.; Zhang, D.; Thiessen, P. A.; Geer, R. C.; Marchler-Bauer, A.; Bryant, S. H. (2014). MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Research, 42(D1), D297–D303. doi:10.1093/nar/gkt1208
3. Dali server. (n.d.). Ekhidna2.Biocenter.helsinki.fi. Retrieved March 14, 2022, from <http://ekhidna2.biocenter.helsinki.fi/dali/>
4. Holm, L.; Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Research, 38(Web Server), W545–W549. doi:10.1093/nar/gkq366