

Vector Alignment Search Tool (VAST): An Information Resource for Protein Structure Comparison

Thomas Madej, Eric W. Sayers, Stephen H. Bryant

National Center for Biotechnology Information (NCBI) ■ National Library of Medicine ■ National Institutes of Health ■ Department of Health and Human Services

VAST is the algorithm used in structure neighboring services provided by NCBI. These include pre-computed similarity relationships among all protein chains and 3D domains in the NCBI 3D structure database, MMDB, which contains structures derived from PDB. VAST is also used for on-the-fly comparisons of new structures against the available database. An acronym for "Vector Alignment Search Tool", VAST rapidly scans candidate structural alignments based on the relative orientations of axial vectors of secondary structure elements. Vector alignments exhibiting surprising similarity, as judged by an objective test statistic computed by VAST, are then refined by an alpha-carbon distance matrix comparison. The collection of pre-computed structure neighbors is available from the NCBI Entrez retrieval service at www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Domains, and it currently contains roughly 90 million 3D alignments. VAST results are summarized using web-based "footprint" graphics to indicate structurally similar regions, and can be displayed as detailed, multiple 3D superpositions using Cn3D, NCBI's molecular graphics viewer. This information is intended to assist biologists in mapping functional sites from one structure to another and to provide a starting point for molecular modeling, using the structure-based alignment tools now available in Cn3D.

Preparing Data for VAST

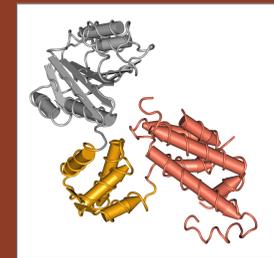


Figure 1. Creating 3D Domains

3D domains of 1A2J chain C. Each protein chain in MMDB is partitioned into domains using a procedure based on inter- vs. intra-domain contact counts. This enhances the sensitivity of the structure comparison algorithm.

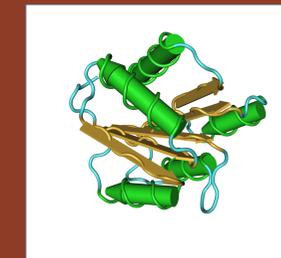


Figure 2. Secondary Structure Representation

Secondary structure representation of 1FQW chain A. Alpha helices are represented by green cylinders, beta strands as tan arrows, and random coil in blue.



Figure 3. Vector Representation

Vector representation of 1FQW chain A. VAST converts structures into sets of vectors, each of which represents a secondary structure element.

How VAST Works

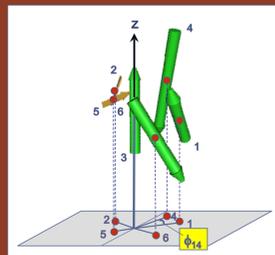


Figure 4. Vector Position About the Z Axis

For both the query and target structures, VAST calculates the midpoint (red circles) for each secondary structure element (SSE). VAST then projects the midpoints onto the XY plane, and then calculates the angles between them.

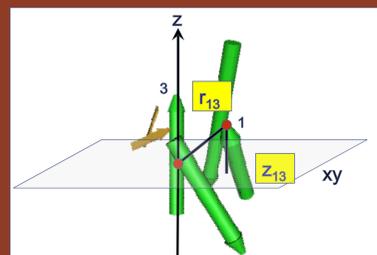


Figure 5. Vector Position From the XY Plane

For each SSE in both the query and target structures, VAST first sets the origin at the midpoint of the SSE and then calculates the distance z to the XY plane and ϕ to the origin.

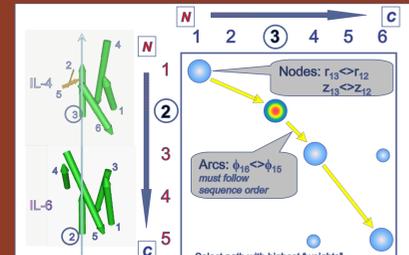


Figure 6. Choosing the Optimum Alignment

To find the optimum alignment, VAST creates a comparison graph between the two vector sets. VAST then weights the graph nodes based on the agreement of ϕ and z between the two sets, and connects these nodes by arcs that are weighted by the agreement of the angles about the Z-axis. VAST selects the path with the greatest weight as the optimum alignment.

PubVAST Statistics

PubVAST is an SQL database used to store VAST results. It is updated monthly and contains the results of pair-wise VAST comparisons between all the protein structures in the RCSB Protein Data Bank. As of January, 2004, it contains the following data:

- more than 44,000 single protein chains
- more than 92,000 domains
- more than 90,000,000 VAST alignments

Viewing VAST Neighbors



Figure 11. Web View of VAST Structure Neighbors

Partial listing of VAST neighbors for 1G83 chain A, the regulatory fragment of Src kinases. As indicated by the Conserved Domain Database (CDD) annotation, this chain consists of two domains, an N-terminal SH3 domain followed by an SH2 domain. The red bars are a "footprint" for the extent of structural similarity, and the display clearly differentiates those representative neighbors that are similar to the SH3 domain, and those that are similar to the SH2 domain. These neighbor representatives are from a non-redundant subset at a "medium" level of redundancy (the default), that corresponds to a BLAST P-value of no more than 10e-40.

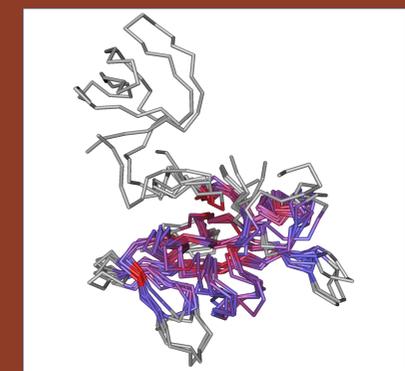


Figure 12. Cn3D View of VAST Structure Neighbors

VAST alignments of several SH3 domains with the SH3 domain of 1G83 chain A. The unaligned part at the top is the SH2 domain of 1G83 A.

Creating and Refining the Vector Alignment

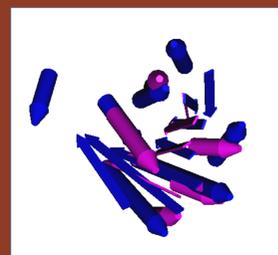


Figure 7. Initial Vector Alignment

Vector alignment between NRC receiver domain (1KRX chain A) and glucose dehydrogenase (1GEE chain A). All of the vectors from the NRC domain align with corresponding vectors from the dehydrogenase. However, the glucose dehydrogenase has two extra alpha helices and two edge beta strands that do not align.

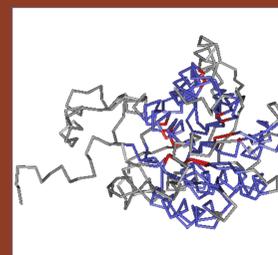


Figure 8. Refined Alignment

Refined alignment between NRC receiver domain and glucose dehydrogenase, based on the initial vector alignment. Non-aligned residues are displayed in gray; aligned residues are in blue/red, with red signifying identical residue types. There are 108 aligned residues with a superposition RMSD of 3.1 Angstroms.

Scoring VAST Alignments

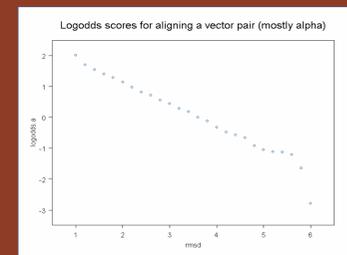


Figure 9. Scoring Alpha Helices

The Logodds scores display the odds of seeing two vector pairs, one pair from each structure, superpose at a given root mean square deviation (RMSD), when the vector pairs are aligned and come from similar structures vs. vector pairs from non-related structures. These pair scores are then used to define the "VAST score", which can be computed for any number of aligned vectors. From the VAST score together with the number of aligned vectors and overall domain sizes and composition, the VAST P-value is calculated, which is an estimate of the statistical significance of the structural similarity.

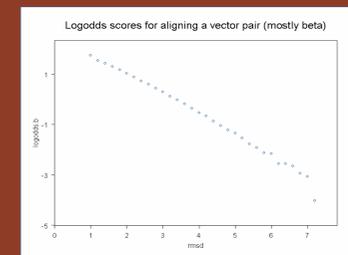


Figure 10. Scoring Beta Strands

Web Addresses

to access VAST pre-computed alignments:

www.ncbi.nlm.nih.gov/Structure

to submit user-generated PDB files to a VAST search:

www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html