CSC 398 - Bioinformatics

Dr. Nakazawa

10th April, 2015

**Vector Alignment Search Tool (VAST) vs. Basic Local Alignment Search Tool (BLAST)**

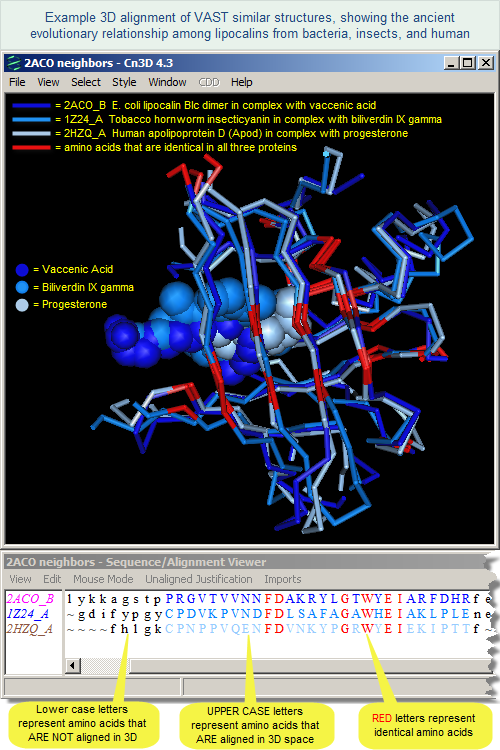
**By Andrés Berejnoi and Phyo Phyo Kyaw Zin**

The computational science of bioinformatics has tremendously advanced the researches in the fields of chemistry, biology and archaeology, since around 1960s, by collecting and distributing information in the form of publicly accessible huge data storage, sorting out sequenced data in interested samples and accurately determining the results by individualized functions to achieve specific objectives of the researches. Bioinformatics tools help visualize nanometer and even smaller scale molecules such as proteins and DNA strands in the forms of letters, interpretable structures on screen in three dimension as well as allow data manipulation to synthesize and design drugs, new proteins for medical purposes. Bioinformatics additionally saves the extensive, tedious workload and reduces human errors by accelerating the alignment process with the use of structure-function predictions based on the enormous database. This paper will compare and contrast two bioinformatics tools expediting specifically in the studies of biochemistry in terms of functionality and how they have been utilized for different goals due to their strengths and weakness. Our selected bioinformatics tools are Vector Alignment Search Tool (VAST) and Basic Local Alignment Search Tool (BLAST).

VAST advanced by National Center for Biotechnology Information (NCBI) classifies protein structures based on their 3D geometric model, without taking into account the similarities or differences present in their sequences. Designed around 1996, VAST has been extensively employed in maintaining MMDB (Molecular Modeling Database) and PDB (Protein Data Bank) to sort out structurally similar proteins for data collection and enabling 3D geometrical comparison of protein configurations from all accessible public data. The algorithm is implemented by vector alignment of secondary structures in identifying important similarities and skipping over irrelevant comparisons among a surfeit of alignment refinements.

The algorithm of VAST employs vector representation to support the secondary structures of proteins in 3D. Streaming SIMD (Single Instruction, Multiple Data) Extensions - SSE is a processor implemented in Gibbs sampling algorithm to compare and contrast with vectors of other secondary structures from selected proteins. Alignment scores, in the form of p-value, are calculated based on the quantity of matched sets of vectors generated by SSE, and vector alignment is finalized through a Monte Carlo procedure. Similar algorithm is applied to calculate the p-value for analogous arrays in BLAST; thereof, both tools have the same level of p-value accuracy for aligning vectors. Correspondingly, VAST is recognized for its ability to diagnose conservative areas and core elements particularly for smaller segments, though high accuracy is not guaranteed. Regardless, different levels of high-conservation accuracy for refinement depending on peculiar intentions of phylogenetic studies is favorable for researches in general.

For the VAST database website, new entries are updated weekly with a fast heuristic model designed for rapid computation of statistics on vector alignments. Nonetheless, the expeditious computational speed of heuristic model comes with one defect of skipping over some possibly intriguing analogues. A screenshot of VAST application demonstrating the structural similarity among lipocalins from bacteria, insects and humans is shown in Fig 1.



**Fig 1: VAST illustrating similarities among lipocalins from bacteria, insects and humans**

Another tool used in protein sequence comparison is BLAST, which stands for Basic Local Alignment Search Tool. Blast is the most commonly used tool for local alignment of sequences. The process of local alignment means that instead of comparing a whole sequence to another, a subset of the query sequence is used to compare against subsets of other sequences in a database. This algorithm is better suited for searches in large databases, given than using global alignment techniques becomes computationally expensive as the amount of different sequences to look through increases.

The way BLAST works is by taking a subset of the query sequence, based on a neighboring score threshold (T), to create a seed that will be used to extend the alignment on both sides of the sequence. The T value for each alignment is calculated using a scoring matrix, which for proteins is typically the BLOSUM 62 matrix. In BLAST, a subsequence is by default 3 characters long. So given a query sequence, the first three characters will be used to form a “word”. Then, more words will be generated by advancing one character in the sequence and taking three more characters. Then each of these words is matched against subsequences in the database to find similarities. The alignment of the two subsequences must have a score higher than the threshold score T (BLAST puts the threshold at 18 by default) in order for the alignment to be extended. Then each position in the sequence against the query is assigned a letter if there is high similarity, a “+” when there is similarity but it is not high, or no symbol, which indicates that there is a non-similar substitution. The term “high similarity” means that the bases are the same or that a substitution for a similar base has occurred. The algorithm also works for nucleotide sequences, using a different scoring matrix. At the end of the search, a report is generated for each alignment found. One important information in the report is the expect value (e) for each alignment. The expect value is used as a probability of how likely it is for the query sequence to be homologous to the aligned sequence, and it is affected by the size of the database used. This means that the same sequence search at a later time can result in a different expect (e) value since the database might have changed. The smaller the expect (e) value, the more likely it is that the aligned sequences are homologous. A value of -5 is often considered sufficient enough when annotating a genome, and a value of -30 is considered to be definite evidence of homology.

The major difference between VAST and BLAST is the different algorithmic approaches implemented in identifying similar protein molecules. VAST software is entirely built on vector alignment for structural comparison while BLAST software mainly resorts to sequence comparison to identify similar protein molecules. Apart from their built-in computational techniques, they share many similarities. First, they share many resourceful databanks such as MMDB and PMD, from where they pick out structures and sequences to compare and compute to get p-value. Secondly, the implemented algorithms to compute the p-value is similar in both VAST and BLAST; consequently, their accuracy in calculating p-value is very high. Third, in comparing huge sequences, they locally align structural fragments other than the whole protracted molecules.

**Bibliography**

1. Thomas Madej, C. J. L. D. Z. P. A. T. R. C. G. A. M.-B. a. S. H. B. MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Research* **2014,** *42,* D297-D303.
2. Marc A. marti-Renom, E. C. I. N. S. a. P. E. B. Structure Comparison and Alignment. In *Structural Bioinformatics,* 2nd ed.; Bourne, J. G. a. P. E., Ed.; John Wiley & Sons, Inc., 2009; pp 395-417.
3. Jean-Francois Gibrat, T. M. a. S. H. B. Surprising similarities in structure comparison. *Sequences and topology* 377-385.