A Boost for Sequence Searching

Extension to BLAST Can Improve its Sensitivity Twofold

With over 47,000 citations, the sequence search method BLAST has been an essential tool in biological research since its development in 1990. By accounting for the influence of sequence context on the mutation probabilities of amino acids, context-specific BLAST (CS-BLAST) achieves two-fold higher sensitivity for distantly related protein sequences, at the same speed and error rate.

**Dr. Johannes Söding, Andreas Biegert,
Gene Center Munich and Center for Integrated Protein Science (CIPSM)**

Homologous proteins, i.e., proteins that have descended from a common ancestor, usually not only have similar sequences but also similar structures and functions. Hence, when two sequences are similar to a degree that cannot be explained by chance, we can assume that the similarity in sequence arose by common descent and that the proteins are therefore likely to be structurally and functionally similar. This principle of "homology-based inference" has proven to be extremely fruitful in clarifying the functions of hundreds of thousands of genes and proteins, due to the simplicity and speed with which searches through large public sequence databases can be done.

By far the most popular program for performing these sequence searches is BLAST (Basic Local Alignment Search Tool) [1, 2]. Although some tools are faster [3, 4] or more sensitive in detecting distant homologs [5, 6], BLAST offers the best combination of speed, sensitivity, and usability. To quantify the similarity between pairs of protein sequences, search methods rely on "substitution matrices". A substitution matrix contains scores for how likely each amino acid is to mutate into any other amino acid. Similar amino acids that are likely to substitute for one another have positive matrix scores, dissimilar ones get negative scores. The total similarity score for two sequences is simply the sum of these substitution matrix scores minus penalties for gaps in the sequence alignment. The maximum-scoring alignment can be calculated with the popular Smith-Waterman algorithm [7]. Fast heuristic filters in BLAST and most other search programs reduce the number of database sequences for which the optimum alignment needs to be calculated.

Context-specific Amino Acid Similarities

A fundamental limitation of substitution matrices and all popular search programs is that the similarity between amino acids is modeled independent of the "sequence context" within which the mutations take place. However, a residue's context can say a lot about what amino acids it is likely to mutate into. A proline, for example, that resides in a natively unfolded region of the polypeptide chain can easily be replaced by small hydrophilic or charged residues, which, together with prolines, occur in very high proportion in natively unfolded regions (fig. 1). But prolines can also be very well conserved in folded parts of proteins, where their unique geometry is often important for the protein's function or stability. A valine within a transmembrane helix region is fairly likely to mutate into a small residue such as alanine or glycine (fig. 1), whereas a valine in the core of a cytosolic folded domain is unlikely to mutate into a small residue. Whether a proline is part of an unfolded or folded region, whether a valine is in a transmembrane helix or in a cytosolic domain is fairly well specified by the context of the surrounding residues.

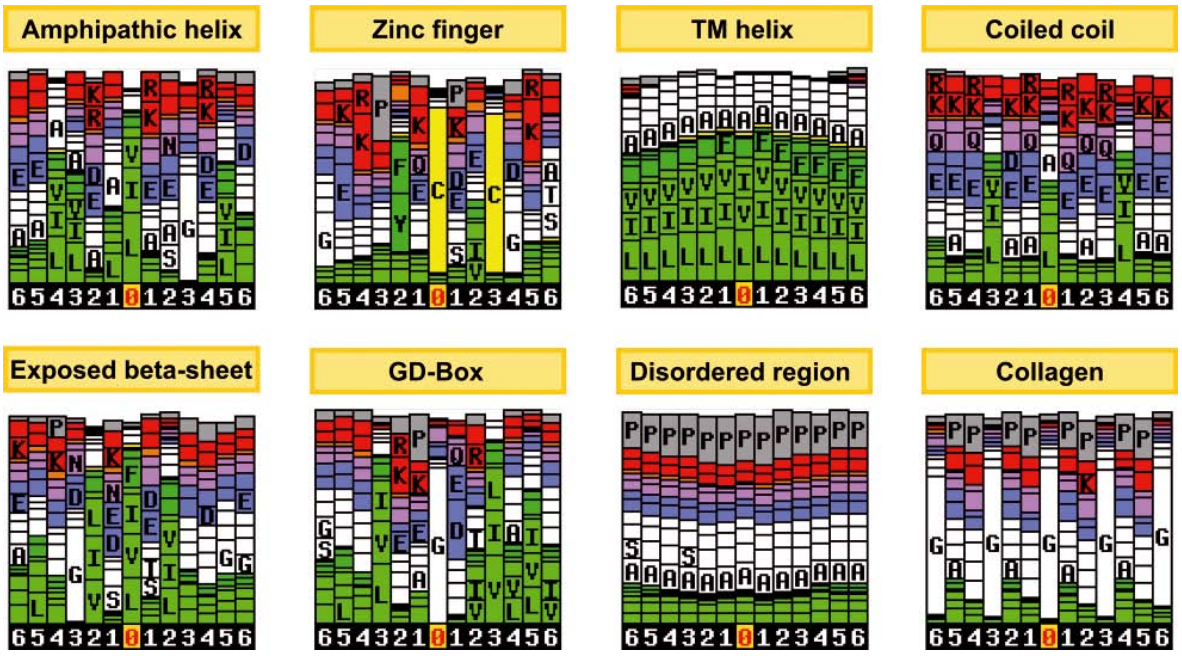

Fig. 1: Examples of amino acid distributions of context profiles describing the sequence contexts learned automatically from a large number of representative sequences. Amino acid colors: green: aliphatic, dark green: aromatic, white: small, red: positive, blue: negative, magenta: polar.

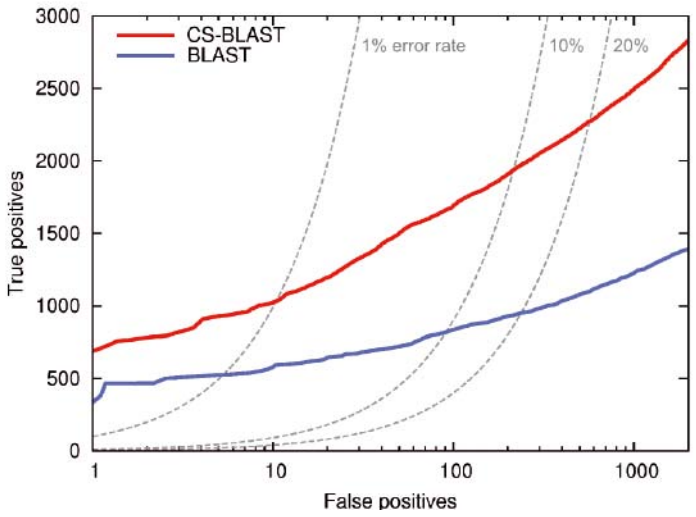

Fig. 2: CS-BLAST detects twice as many remotely homologous sequences at specified error rate than BLAST. The plot is based on a benchmark of sequences with known structure for which the homology relations are well known. It shows the number of homologous sequence pairs ("true positives") detected at a given number of falsely detected non-homologous sequence pairs ("false positives"). To exclude testing on trivial cases, no sequence pair in the test set has more than 20% sequence identity.

Context-specific BLAST

Context-specific BLAST is an extension of BLAST that uses context-specific substitution scores to calculate the similarity between protein sequences [8]. By making use of context information, CS-BLAST is able to find twice as many remotely homologous sequences at specified error rate as BLAST (fig. 2). Likewise, the quality of the alignments it produces is significantly higher, particularly for the most difficult alignments [8].

CS-BLAST is a wrapper around BLAST that calculates context-specific amino acid similarities for the query sequence. These are written into a position-specific scoring matrix, also called "sequence profile". A profile contains for every position in the query sequence a column with the similarity scores of the 20 amino acids. CS-BLAST then jump-starts BLAST with this profile. Hence, all heuristic optimizations of BLAST also benefit CS-BLAST. The actual search, which takes the same time whether started with a single sequence (as in BLAST) or with a sequence profile (as in CS-BLAST), typically takes a couple of minutes, much longer than the couple of seconds for calculating the profile with the context-specific amino acid similarities. Therefore, CS-BLAST runs at roughly the same speed as BLAST despite its improved sensitivity. Switching from BLAST to CS-BLAST is simple: Once BLAST is installed; CS-BLAST can be run directly, producing the same output format.

To calculate the context-specific amino acid similarities for a query sequence, CS-BLAST cuts it into overlapping windows of 13 residues length. Each window is compared with a library of 4,000 context profiles (fig. 1). The context-specific amino acid scores for the central residue of each query sequence window are obtained by combining the similarity scores in the central columns of the 4,000 context profiles.

Each context profile contributes in proportion to its similarity with the query sequence window. As an example, the amino acid similarity scores for a valine residue within a transmembrane helix will be contributed mainly by those context profiles which describe transmembrane helices.

Learning the Contexts

The library of 4,000 context profiles is trained automatically from a representative set of 50,000 sequences randomly picked from the public databases. The sequences were cut into overlapping windows of 13 residues and clustered into the 4,000 context profiles. Similar sequence windows tend to end up in the same context profiles. Sequence contexts that appear frequently in the database are described with finer detail in this context library (i.e., with more profiles) than rare contexts. The choices of the number and length of the context profiles result from a trade-off between sensitivity and time to calculate the amino acid similarities.

All Bio-sequences Have Contexts

The new paradigm of sequence specificity is very general and can be applied not only to sequence-sequence comparison, but to all aspects of sequence comparison or molecular evolution. As an illustration, we have developed a context-specific version of the sensitive and widely used position-specific iterated BLAST (PSI-BLAST) [2]. Here, a sequence database is iteratively searched with a sequence profile and significant sequence matches are included into the profile for the subsequent iteration. CSI-BLAST, our context-specific version of PSI-BLAST, finds as many homologous proteins after two search iterations as PSI-BLAST after five iterations [8].

Next, we will extend the paradigm of context-specificity to the alignment of genomic sequences. Mutations in cis-regulatory elements, such as transcription factor binding sites, may be more important for evolutionary changes than mutations in coding regions [9] and a number of genetic diseases is linked with mutations in cis-regulatory elements [10]. But whereas the majority of the human genes have been identified and assigned partial functions, only a tiny fraction of the cis-regulatory elements and their functions is known so far. With substantially improved genomic alignments, we hope to be better able to detect functional elements on a genome-wide scale through their conservation among related species.

Free Web Servers

We have set up a server for testing CS-BLAST at http://toolkit.lmb.uni-muenchen.de/cs_blast. The Swiss Institute for Bioinformatics will offer free access to CS-BLAST searches on their EMBnet server: http://www.ch.embnet.org/software/aBLAST.html. The CS-BLAST binaries can be downloaded from ftp://toolkit.lmb.uni-muenchen.de/csblast/. They are free for academic use. For commercial use, please contact us for a license.

References

[1] Altschul S.F. *et al.:* J. Mol. Biol. 215, 403–410 (1990)
[2] Altschul S.F. *et al.:* Nucl. Acids Res. 25, 3389–3402 (1997)
[3] Ma B. *et al.:* Bioinformatics 18(3), 440–445 (2002)
[4] Morgulis A. *et al.:* Bioinformatics 24(16), 1757–1764 (2008)
[5] Pearson W.R. and Lipman D.J.: Proc Natl Acad Sci USA 85, 2444–2448 (1988)
[6] Eddy S.R.: Bioinformatics 14, 755–763 (1998)
[7] Smith T.F. *et al.:* J Mol Evol. 18, 38–46 (1981)
[8] Biegert A. and Söding: J., Proc Natl Acad Sci USA 106, 3770–3775 (2009)
[9] de Vooght K.M.K.: Clinical Chemistry 55, 698–708 (2009)
[10] Wray G.A.: Nat Rev Genet. 8, 206–216 (2007)

Contact:

Andreas Biegert
Dr. Johannes Söding
Gene Center Munich and Center for Integrated Protein Science (CIPSM)
Ludwig-Maximilians-Universtät München
Munich, Germany
soeding@lmb.uni-muenchen.de