33. BlastXtract: a web-based tool for managing and visualizing results from large translated BLAST and FastA searches

Marcus Claesson

Alimentary Pharmabiotic Centre and Department of Microbiology, National University of Ireland, Cork, Ireland

Searches of translated, un-annotated genomic DNA against protein databases is a useful early-stage method for discovering encoded protein homologues, but generates huge amounts of data that quickly becomes impregnable. BlastXtract is a web-based tool for managing and visualizing results from large translated BLAST and FastA searches. It combines the speed and storage benefits of RDMS with an easy-to-use graphical navigation map, and greatly facilitates the early exploration of genomic sequence.

The ever increasing amount of genomic data being produced brings the need for tools that can easily manage and visualize results from sequence similarity programs. The most widely used tool for searching genomic sequence for homologues is BLAST (Altschul et al., 1997), and performing translated searches such as BLASTX agains protein databases can predict the presence of many putative genes (Robison et al. 1994). This is a relatively quick and simple method to get an overview of the contents of a genome, without running gene prediction programs and subsequent sequence analysis of putative proteins. However, searches of large genomic query sequences generate vast amounts of output data, which quickly becomes unmanageable and impossible to browse through. Hence, a need exists to devise a bioinformatics too that reliably stores all produced data, and is also capable of performing fast searches and orderly retrieval of particular results, visualized by a user friendly graphical interface.

Another important issue during assembly and early annotation of genomic DNA sequences is to identify sequencing and assembly errors, which are frequent in especially low quality or draft genome sequences and can cause frameshifts. Translated BLASTX searches produce two High Scoring Pair (HSP) alignments around a frameshift site and these can be visualized graphically to aid detection. Therefore, in order to combine the benefits of an easy-to-use graphical interface with the speed and high storage capacity of RDBMS, BlastXtract was developed. This novel web-based tool allows translated BLAST results to be uploaded into a database where it can be further queried and visualized through an intuitive and clickable navigation map.

In addition to managing BLAST outputs, BlastXtract is also able to include translated results from the FastA program (Pearson & Lipman, 1988), which utilizes a slower but more sensitive pairwise alignment algorithm. Unlike BLAST FastA produces only one alignment per hit and frameshifts are consequently colour coded to aid detection The combination of displaying BLAST and FastA searches with their alignments side-by-side is more informative and increases the likelihood of making correct gene predictions.

IMPLEMENTATION AND METHODS

A new BlastXtract session commences with the user uploading a standard output file of either BLASTX or FastX/FastY results into the database. The choice of parameters of the initial search is left to the user, but for a high level of detail and statistical significance it is recommended to allow as many relevant hits as possible and to divide longer query sequences into smaller ones. In the case of the latter, BlastXtract's query offset function allows the pieces to be `stitched' back together to mimic a search of the full sequence. Once the result file has been uploaded, with an optional data description, into a database table it can be browsed and explored further. In addition to the translated DNA-to-protein searches also DNA-to-DNA output data can be processed. However, the translated searches are more useful for annotation purposes, since protein sequences have a higher degree of conservation compared to DNA and their database entries usually are more informative. The user can choose to look at all possible hits within a specified query sequence range or limit the search for hits with certain words or accession numbers in its description. Hits that overlap can be filtered out, which is a very useful way to get an overview of only the best hits for each position. Also E-value thresholds can be set. The way to display the hits can be either graphical or non-graphical. The non- graphical display shows all the values of the HSPs in a table and gives the option of showing the full protein alignment for each case. The accession numbers in the hit description are hyper-linked to the SRS and NCBI web server. Every HSP also has two bar graphs which illustrate the relative positions within the total query sequence and the chosen range. They also indicate where in the sequence the start and stop would be if the alignment was complete. The graphical display visualizes an overview of the requested hits in the specified sequence range and shows the values of every hit when scrolled over. HSPs that belong to the same hits are intertwined with dotted lines. This view also works as a navigation map where more detailed information can be obtained as in the non-graphical display, by clicking on the hits. A colour coding scheme indicates in which frame the HSPs are found and how high score it has.

BlastXtract has been an important tool in the annotation process of two prokaryotic genomes, Lactobacillus salivarius UCC118 (low-GC content and high sequence read coverage) and Bifidobacterium breve UCC2003 (high-GC content and low coverage), where the detection of around 400 frameshifts was greatly facilitated.

BlastXtract was written in Perl and uses Bioperl modules for parsing the output data and producing the graphical objects. The web interface runs under Apache web server and is built using the Perl CGI module and JavaScript. The Perl DBI module enables communication with a RDBMS, which can be either MySQL or PostgreSQL. On the server side BlastXtract needs to be installed on a Linux/UNIX system with the required perl modules and database systems, but from the client side only a standard web browser is needed.

ACKNOWLEDGMENTS

Thanks go to Rory Mullane for his appreciated help with the web design, and to the UCC Bioinformatics lab for their supportive suggestions. This work was supported by a Food Institutional Research Measure grant (01/R&D/C/159) through the Department of Agriculture & Food under the National Development Plan 2000-2006 and the Science Foundation Ireland funded Alimentary Pharmabiotic Centre.

REFERENCES

Altschul, S. Madden, T. et al. (1997) Gapped BLAST and PSI_BLAST: a new generation of protein database programs. Nucleic Acids Research., 25, 2289-3402.

Pearson, W. R. Lipman, D. J. (1988) Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the USA., 4, 2444-2448.

Robison, K. Gilbert, W. Church, G.M. (1994) Large scale bacterial gene discovery by similarity search. Nature Genetics 7 (2), 205-214.