A Data Mining Tool for Frame-Shift Smith-Waterman Outputs

Keith E. Joho, David J. Chang, Roche Bioscience, Palo Alto, California and
Albert F. Shpuntoff, AFS Informatics, Mountain View, CA

Keith.Joho@roche.com, David-J.Chang@roche.com,
afs@world.std.com

This is a first presentation of a data mining tool developed at Roche Bioscience for sifting through the mounds of outputs that are often produced when a Time Logic Decypher system is used to perform frame-shift tolerant Smith-Waterman searches.

In the first version of the system, we typically begin with a collection of many (related) protein sequences, which collectively represent an interesting family. The Smith-Waterman algorithm is then used to search a nucleic database, using scoring criteria that are designed to find even relatively weak homologies. By setting the penalties and thresholds low and requesting large numbers of outputs, a very large collection of outputs can be guaranteed. The sequences, which are given high P-score, represent an obvious avenue of further study, but often researchers are interested in finding sequences, which are of interest for other reasons. The browsing tool was designed to digest the mounds of outputs and to provide the user with the ability to rapidly identify sequences occurring in the output collection by providing a fairly comprehensive set of parameters for parsing the outputs.

On first importation into the browser’s system, a Decypher output is scanned for the key values generated by each sequence in the output and indexed. Simple flat files are used to contain the values used for further selections, which includes raw score, P-score, percent identity, percent similarity, presence of annotation as public-data sequence, presence of stop-codons in the protein representation of the aligned sequence, and length of the target sequence segment present in the alignment. In addition, as the indexes are prepared, each sequence is annotated with the number of times that the sequence occurs as a hit in the collection of outputs that are being digested.

An initial menu is presented to the user, announcing the number of hits that are available for inspection. By choosing various properties that must be met, the user restricts the population of hits to sequences that are potentially of interest. A fairly common set of restrictions requests first the exclusion of scores with near-perfect P-scores, then restricting attention to sequences which occur sufficiently often as to be interesting. At any point in the browse/restrict process, the user can display the (alignments produced for all) active sequences, and click on each sequence to place it into a “selected” category for further study. Also, at any point in the process, the user may re-activate all of the sequences present in the output, sorted either by acquisition number or by P-value.

At subsequent sessions, additional outputs may be added to an existing project, and additional sequences may be placed in the “selected” category. The list of “selected” sequences is kept in a text file, suitable for use in subsequent command procedures.

During the initial trials of this browser, many interesting annotations have been developed for both public and private data. We have been able to easily identify very distantly related family members that only share 20-25% identity with the query sequence.