CBSU Software

Jaroslaw Pillardy
Qi Sun

Computational Biology Service Unit
Cornell Theory Center

Hardware infrastructure

CBSU has 128-CPU (64 node) cluster for research, each CPU is Pentium III 1GHz, with 2GB RAM memory per node.

There are other clusters at the Theory Center:

V1: 256 CPUs (64 nodes) – Pentium III Xeon 500MHz
V+: 128 CPUs (64 nodes) – Pentium III 733MHz

They may be also used in research.

We are therefore focused on developing software for efficient use of our parallel environment.

CBSU Software

CBSU Web Page

RepeatFinder

CBSU BLAST

Parallel BLAST

Parallel LOOPP

CBSU Web Page

- Active toolbar
- Software depository
- Educational resources
- Web-based software
- Interactive bioinformatics web resources database

RepeatFinder

http://ser-loopp.tc.cornell.edu/cbsu/repeatfinder.htm

This program identifies small repetitive sequences scattered in the genome. Possible repeat sizes are from about 15 to few hundred bp.

- It is highly configurable and flexible
- It is capable of dealing with large sequences
- It is not only identifying repeats, but it also clusters them and estimates statistical significance
- The resulting repeats are also clustered together based on partial sequence overlaps and their statistical significance
- It provides easy-to-read, HTML-based output, as well as spreadsheet data files

RepeatFinder

http://ser-loopp.tc.cornell.edu/cbsu/repeatfinder.htm

What is a repeat?

Most probable sequence in a group of highly overlapping sequences scattered throughout the genome.

Accounts for variability but is very sensitive on clustering algorithm used. Non-trivial clustering problem, but it leads to a very flexible algorithm.

RepeatFinder: algorithm

http://ser-loopp.tc.cornell.edu/cbsu/repeatfinder.htm

- The input sequence is iteratively blasted against itself after it is divided into multiple overlapping blocks
- Matrix of relationships for all putative repeats of appropriate length is calculated
- The repeats are clustered into groups by analyzing the relationship trees, minimal tree algorithm with end-point deviation metric is used.
- Optimal multiple sequence alignment in each group is calculated and variable ands are trimmed.
- E-value and variability in each group are calculated, too small or too variable groups are eliminated and most probable sequence is calculated
- Most probable sequence from each group is blasted against the input sequence to check for missed matches.
- The groups are clustered together based on statistical significance or inter-group overlaps

RepeatFinder: output

http://ser-loopp.tc.cornell.edu/cbsu/repeatfinder.htm

- Information is presented as plots as well as as numbers.
- All the plots are clickable; a message box will present numerical value(s) of the field being clicked. In the case of distribution plot it will show the boundaries of the segment being clicked (size of the segment is reported above the plot).
- RMSD of the repeat's length is calculated as root-mean-square deviation from the average length.
- Composition of a group is shown in a shaded table. Shade of the element is proportional to its numerical value. The first four rows show nucleic acid occupancies (A, C, G and T), the fifth one shows gaps. The sixth row 'P' reports the probabilities of a given field to be occupied by any nucleic acid (not being a gap), and the last row 'V' shows variability.
- Variability is calculated as a probability of a given position being occupied by something else than the most probable occupant (0.0 - only one kind of occupancy, 1.0 - flat distribution).

RepeatFinder: output

http://ser-loopp.tc.cornell.edu/cbsu/repeatfinder.htm

- *Cumulative E-value* of a group is the probability of finding such a group of sequences (assuming the member's probabilities of occurrence are not correlated to each other) in the whole sequence at random.

- *True E-value* of a group is the cumulative E-value corrected for misalignments and inaccuracies within a group. If all the sequences in a group are identical it is equal to cumulative E-value.

- *True composition-corrected E-value* of a group is the true E-value where the score for each sequence is corrected according to a local sequence composition (size of a local window is given by *local_win* parameter).

- *E-value for a single element* of a group is the standard E-value for an average-length element of a group. Given for comparison only.

- *E-value for all the elements merged* is the standard E-value for a single sequence produced by merging all the group elements together. Given for comparison only.

CBSU BLAST

Iterative masking parallel BLAST implemented in the CBSU.

- BLAST is carried out in iterative manner where previous results are masked for next iterations. It assures that all possible hits will be reported, even when non-redundant databases are used, or particularly popular pattern is encountered.
- Results are presented in a hierarchical way, where similar matches are presented together, and only the best representative of a group is shown in the top document. Access to all matches is provided through hyperlinks.
- It is fully parallelized, and therefore it is capable of handling large sets of sequences fast.

CBSU BLAST

CBSU BLAST is especially useful when your data is possibly contaminated by vector sequence(s) or non-redundant database is used (htg). Results from standard BLAST may not only miss important hits in this case, but also may be almost impossible to read.

CBSU BLAST is designed for interactive human use. For database creation or high-throughput processing parallel BLAST at CBSU should be used.