

CLOBB.pl - A program for clustering sequences on the basis of BLAST similarity

SYNOPSIS

CLOBB.pl <cluster_id>

DESCRIPTION

This program takes a set of DNA sequences and clusters them into groups which putatively derive from the same gene. In order to operate, the user must have NCBI's blastall executable in their path. The output is a blastable fasta file named <cluster_id>EST which contains a list of the sequences and a cluster identifier <cluster_id>EST00001 -> <cluster_id>EST99999. A number of other files and directories are also created.

HOW IT WORKS

The program works by performing a BLAST for each sequence against the growing cluster database. The BLAST output is then examined for High Scoring Pairs (HSPs) from the BLAST output

which demonstrate near identical regions of sequence similarity (>=95% sequence identity over >30 bases - these figures are modifiable in the source code to increase/decrease stringency) between the target sequence and the growing cluster database. These matches are classified as type I matches.

The next stage of the process goes through the list of type I matches to identify any problems associated with the match. This is achieved by parsing the beginning and end positions of the query and subject sequences from the blast output. If these positions overlap beyond the HSP by more than 30 bases (i.e. the HSP does not extend through the full overlap of the sequences), a further check is performed to ensure that this is not due to the presence of poor quality sequence (determined by the number## of bases assigned `N' in the overlap regions). Type I matches which do not have high quality overlaps of more than 30 bases beyond the HSP are designated as type II matches. Other type I matches which possess high quality overlaps of greater than 30 bases which are not part of a HSP are designated as type III matches.

The next stage of cluster assignment then involves checking through the lists of type II and type III matches to ensure that no conflicts arise. Given a cluster in which some members are type II matches, if there are other members of the same cluster which have been designated as type III matches, then this indicates that the query sequence matches some but not all members of a cluster and is therefore assigned a new cluster number. The inclusion of this feature in the algorithm prevents the rapid expansion of chimeric clusters and can result in a splitting of related sequences into many different (related) clusters. However, when such events occur, the program catalogues the clusters involved, identifying them as `similar to' the type III match for subsequent post-process analysis (typically performed by manual curation).

Another complication occurs when two or more type II matches arise from different clusters. Firstly the blast output is reanalysed to determine whether the HSPs of the matches occur in overlapping regions. If they do not, the query effectively links the clusters and they are merged into the cluster with the lowest index - a separate note is recorded to indicate that such a merge operation has occurred. If they do overlap, this may indicate that they are either alternatively spliced variants of one gene or closely related members of a gene family, and the query sequence is assigned the cluster number of the type II match with which it had the highest blast score, providing that said cluster did not contain a type III match, and an annotation added to indicate that these clusters## may be members of a `Supercluster'. Once the query sequence has been assigned to a cluster, it is added to the growing cluster database which is then reformatted to allow the next search.

Created files/Directories

*** supercluster**

This file contains a list of clusters, members of which were found to be similar to each other, however, other members of the cluster were found to be bad matches

*** merge**

This file contains a list of clusters which have been merged to form a single cluster. This occurs when a sequence matches well with two clusters which do not overlap.

*** master**

A file containing a comma delimited list of the sequences.

*** sequences_done**

A directory listing all sequences which have been processed by CLOBB. Two sequences are created for each original, the original sequence is renamed as <sequence>.old, the new sequence file has the assigned cluster id annotated to its header.

*** OUT**

A directory containing a logfile detailing the CLOBB process for each sequence.

AUTHOR

John Parkinson (john.parkinson@ed.ac.uk)

COPYRIGHT

This program is free software; you can redistribute it and/or modify it under the same terms as Perl itself.

SEE ALSO

perl(1).