Poster H-40

GFMerge - A method for merging of gene predictions

Authors:
Sebastian Spiegler (*Technical University Ilmenau, Germany; Wellcome Trust Sanger Institute, Cambridge, UK*)
Marie-Adele Rajandream (*Wellcome Trust Sanger Institute, Cambridge, UK*)
Neil Hall (*The Institute for Genomic Research (TIGR), Maryland, USA*)
Kim Rutherford (*Genetics Department, Cambridge University, UK*)
Arnaud Kerhornou (*Genome Bioinformatics Research Group, Barcelona, Spain*)

Short Abstract: Genome data from large-scale sequencing is primarily scanned for possible genes using gene prediction programs. Results often contradict each other in structure and location. "GFMerge" calculates a single consensus of non-overlapping high-quality predictions. Within an internal hierarchical scoring system it uses additional similarity information, considers gene structure and evaluates each gene prediction program.

Long Abstract:

Motivation

Large-scale sequencing generates tremendous amounts of raw genome data which have to be examined for possible genes. First examinations are done by the use of gene prediction programs which can be similarity or statistically based. Overlapping gene predictions often differ in structure and therefore have to be manually inspected by biologists.

Problem statement

In order to keep up with data generation, the process of manual inspection should be automated. Furthermore, it may be possible to improve the overall quality of automatically predicted genes for one sequence by combining only those which are of high quality.

Approach

"GFMerge" was developed to calculate a consensus containing only non-overlapping high-quality gene models.

The input for the program is an arbitrary number of gene predictions files from gene prediction programs for a specified sequence as well as protein similarity data (BlastX) and similarity information derived from mRNAs (EST). Moreover, an independent evaluation data set is used for the evaluation of each gene prediction program. All input files must be in EMBL format.

During a pre-processing step, GFMerge clusters the sequence into non-overlapping regions which contain overlapping gene predictions themselves. This is done by a recursive clustering algorithm. Comparisons and selection considerations are only made among

predictions within one region.

The evaluation process itself consists of a hierarchy of six analysis steps. In each step a scoring function rates all gene predictions. Afterwards a recursive algorithm calculates the high-scoring path within each region by maximising the accumulation of single scores of non-overlapping gene predictions. Only those gene predictions which are located along the path are kept, others are assumed to by of lower quality and are discarded.

The hierarchy of analysis steps follows the approach that best evidence is used first. Lower quality evidence for choosing between overlapping predictions is only used if no other information is available. During the first step, internal splice sites are confirmed using EST data. Then, exons as a whole are confirmed by their overlap with ESTs. Then similarities to proteins are considered by comparing remaining gene predictions against protein sequences (using BLASTX). During the next two steps, gene predictions with long exons and introns are favoured in order to keep as much coding information and to cover as much sequence space as possible. At this level remaining overlaps should only consist of predictions of equal or similar structure. In a last round, predictions of better gene prediction programs are preferred over less accurate programs. Gene prediction programs are evaluated in advance by comparing their results to the manual reviewed evaluation data set.

Results and conclusion

GFMerge is a modular, Java-based application. Benchmark tests on parts of the Dictyostelium discoideum chromosome 6 demonstrated that the overall quality of predictions from different gene prediction programs could be improved by merging predictions to a single conclusion only containing non-overlapping high-quality predictions. Furthermore, heavy-load tests on sets up to 15 mega bases showed that processing time does not exceed polynomial complexity.

GFMerge is a hands-on solution to automate manual inspection of gene predictions coming from different gene prediction programs. It demonstrated its abilities during the Dictyostelium discoideum Genome Project ("The genome of the social amoeba Dictyostelium discoideum." Nature 435, 43-57, 2005) for which it was primarily developed. Currently, GFMerge is being used as part of the automatic analysis for the Eimeria tenella Genome Project (http://www.sanger.ac.uk/Projects/E_tenella/).