New gene prediction method capitalizes on multiple genomes

December 20 2007

Researchers at Stanford University report in the online open access journal, *Genome Biology*, a new approach to computationally predicting the locations and structures of protein-coding genes in a genome. Gene finding remains an important problem in biology as scientists are still far from fully mapping the set of human genes.

Furthermore, gene maps for other vertebrates, including important model organisms such as mouse, are much more incomplete than the human annotation. The new technique, known as CONTRAST (CONditionally TRAined Search for Transcripts), works by comparing a genome of interest to the genomes of several related species.

CONTRAST exploits the fact that the functional role protein-coding genes play a specific part within a cell and are therefore subjected to characteristic evolutionary pressures. For example, mutations that alter an important part of a protein's structure are likely to be deleterious and thus selected against. On the other hand, mutations that preserve a protein's amino acid sequence are normally well tolerated. Thus, protein-coding genes can be identified by searching a genome for regions that show evidence such patterns of selection. However, learning to recognize such patterns when more than two species are compared has proved difficult.

Previous systems for gene prediction were able to effectively make use of one additional 'informant' genome. For example, when searching for human genes, taking into account information from the mouse genome

led to a substantial increase in accuracy. But, no system was able to leverage additional informant genomes to improve upon state-of-the-art performance using mouse alone, although it was expected that adding informants would make patterns of selection clearer.

CONTRAST solves this problem by learning to recognize the signature of protein-coding gene selection in a fundamentally different way from previous approaches. Instead of constructing a model of sequence evolution, CONTRAST directly 'learns' which features of a genomic alignment are most useful for recognizing genes. This approach leads to overall higher levels of accuracy and is able to extract useful information from several informant sequences.

In a test on the human genome, CONTRAST exactly predicted the full structure of 59% of the genes in the test set, compared with the previous best result of 36%. Its exact exon sensitivity of 93%, compared with a previous best of 84%, translates into many thousands of exons correctly predicted by CONTRAST but missed by previous methods. Importantly, CONTRAST's accuracy using a combination of eleven informant genomes was significantly higher than its accuracy using any single informant. The substantial advance in predictive accuracy represented by CONTRAST will further efforts to complete protein-coding gene maps for human and other organisms.

Further information about existing gene-prediction methods and the advance CONTRAST brings to the field can be found in a minireview by Paul Flicek, which accompanies the article by Batzoglou and colleagues.

Source: BioMed Central