Researchers develop new self-training gene prediction program for fungi

September 29 2008

Researchers at the Georgia Institute of Technology have developed a computer program that trains itself to predict genes in the DNA sequences of fungi.

Fungi – which range from yeast to mushrooms – are important for industry and human health, so understanding the recently sequenced fungal genomes can help in developing and producing critical pharmaceuticals. Gene prediction can also help to identify potential targets for therapeutic intervention and vaccination against pathogenic fungi.

"While we previously showed that our unsupervised training program worked well to predict genes in many eukaryotes, it didn't work as well for various fungal genomes that carry a significant part of the information that facilitates accurate gene prediction in locations called branch point sites," said Mark Borodovsky, director of Georgia Tech's Center for Bioinformatics and Computational Genomics.

Branch point sites are located inside introns, which are non-coding regions of DNA located between genetic-code carrying regions called exons.

"Previously during the process of predicting the exon-intron structure of eukaryotic genes, we didn't search for branch point sites, but doing so in the new program helps to better delineate intron regions inside fungal genes," added Borodovsky, who is also a Regents' Professor in the

Coulter Department of Biomedical Engineering and the Computational Science and Engineering Division of the College of Computing.

Borodovsky and his colleagues expanded the eukaryotic genome self-training software program they developed in 2005 to address the issue that fungal genes are more complex than other eukaryotes. The research team included graduate student Vardges Ter-Hovhannisyan, Wallace H. Coulter Department of Biomedical Engineering research scientist Alexandre Lomsadze and School of Biology professor Yury Chernoff.

Details of the new program, called GeneMark.hmm-ES (BP), are available online in the journal *Genome Research* and will be included in the journal's December print edition. The software will also be freely available for academic researchers.

Borodovsky developed the first version of GeneMark in 1993. In 1995, this program was used to find genes in the first completely sequenced genomes of bacteria and archea. The research team then developed self-training versions of the gene finding program for prokaryotic (organisms that lack a cell nucleus) and eukaryotic (organisms that contain a cell nucleus) genomes in 2001 and 2005, respectively. Development of these programs has been supported by the National Institutes of Health.

Unlike other programs that require a pre-determined training set along with the genome sequence, GeneMark.hmm-ES (BP) only requires the genome sequence. The program is able to iteratively identify the correct algorithm parameters from the anonymous sequence. The program uses a probabilistic mathematical model called the Hidden Markov Model to pinpoint the boundaries between coding sequences (exons) and non-coding sequences (introns and intergenic regions).

Most introns start from the dinucleotide guanine-thymine (abbreviated GT) and end with the dinucleotide adenine-guanine (abbreviated AG).

However, finding these dinucleotides is not sufficient to signal the presence of an intron. Several nucleotides that surround GT and AG are also important, but the similarity of the pattern is not deterministic. Locating the branch site – which is nine nucleotides in length, almost always contains an adenine and is located 20-50 bases upstream of the acceptor site – helps to accurately identify an intron.

An initial run of the program with a reduced model containing heuristically defined parameters breaks the sequence into coding and non-coding regions. With this information, the researchers apply machine-learning techniques to refine the parameters of the recognition algorithm with respect to the specific patterns found in the newly identified protein-coding and non-coding sequences as well as the border sites.

The prediction and training steps are repeated, each time detecting a larger set of true coding and non-coding sequences that are used to further improve the model employed in statistical pattern recognition. When the new sequence breakdown coincides with the previous one, the researchers record their final set of predicted genes.

To test the algorithm, the researchers selected 16 fungal species from the phyla Ascomycota,
Basidiomycota and Zygomycota and compiled sets of genome sequences containing previously validated genes. The species spanned large evolutional distances and exhibited significant variability in genome size, gene number and average number of introns per gene. The results showed that by including branch site information in the model, the researchers could more accurately predict exon-intron structures of fungal genes.

"The enhanced program predicted fungal genes with higher accuracy than either the original self-training algorithm or known algorithms with

supervised training," noted Borodovsky. "And because we didn't need any additional training information for our program, the sequencing teams could immediately proceed with gene annotation right after the genomic sequence was in hand without spending time and effort to extract a set of validated genes necessary for estimating parameters of traditional algorithms."

Researchers at the U.S. Department of Energy Joint Genome Institute and the Broad Institute of the Massachusetts Institute of Technology and Harvard University have already realized the advantages of the new algorithm. They have already used the new program to annotate about 20 novel fungal genomes. In addition, hundreds of fungal genome sequencing projects currently in progress should benefit from the new method as well, according to Borodovsky.

With the fungal software completed, Borodovsky and his team are already looking to expand their gene prediction algorithms to accurately interpret even more complex eukaryotic genomes.

"There are genome sequencing projects where large repeat populations, a significant number of pseudogenes or substantial sequence inhomogeneity hamper ab initio gene prediction and we're ready to tackle them next," added Borodovsky.

Source: Georgia Institute of Technology