Application of ESTs mapping to improve gene prediction methods

L. Milanesi, I. Rogozin, R. Rizzi

CNR-ITBA - Via Fratelli Cervi 93, 20090 Segrate (Milano)

Prediction of protein-coding genes in newly sequenced DNA becomes very important in large genome sequencing projects. These problems are complicated due to exon-intron of the eukaryotic genes. Currently existing collections of expressed sequence tags (ESTs) are very large and thus very useful for gene mapping. Gene identification in the newly-discovered DNA sequences is an important problem in current molecular biology studies. A number of programs have been developed for predicting the protein coding genes. The most common approach is based on the combination of the potential functional signals with global statistical properties of protein coding regions. Another approach for gene structure prediction is based on the homology detection throughout the databases of nucleotide or amino acid sequences. By using the information available on homologous protein sequences, it is possible to significantly improve the accuracy of gene structure prediction. Currently existing collections of expressed sequence tags (ESTs) are very large and can be very useful for gene mapping. Homology searches against the EST Division of GenBank (dbEST) and Unigene database can be used for this purpose. ESTs (Expressed Sequence Tags) offer a rapid route to gene identification (Adams, et al, 1991, Adams , et al, 1992), analysis of expression and regulation data, and can highlight multigene family diversity and gene alternative splicing). EST matches may identify more than half of the known human genes (Hillier et al, 1996). The price of the high-volume and high-throughput nature of the data, however, is that ESTs contain high error rates (Aaronson, et al 1996), do not have a defined protein product, are not well annotated and present only a raw substrate for sequence matching.

The ESTMAP system involves the following procedures:

1) Repeat masking. The repeated elements (for example, the human Alu elements) can be automatically masked in a query sequence before the homology search. Homology searches against the collection of repeated element (Jurka et al., 1992) are used for repeats detection. We implemented a program REPEAT for that purpose. A censored sequence (with 'N's instead of repeated elements) is automatically produced by REPEAT.

2) Homology searches. BLASTN (Altschul et al. 1990) is used for homology searches of the censored query sequence against the EST Division of GenBank (dbEST) and the Unigene database of sequences (www. ncbi.nlm.nih.gov) This step is most time-consuming since these EST datsets are very large.

3) EST mapping. The BLASTN output is used as input information by a EST_GENE program. Information about an EST sequence is used only when the similarity between the EST sequence and the query sequence is greater then 95%. The module EST_GENE is also able to predict the introns in DNA comparing ESTs and a query sequence based on the alignment method suggested by Huang (1994) (a linear-space divide-and-conquer strategy). The GT/AG splicing sites rule is used by EST_GENE, however non-canonical splicing signals (Milanesi and Rogozin, 1998) can also be predicted in cases of unambiguous alignment.

4) Output of results. The graphical visualization of the results is particularly important for the analysis of alternative splicing in a query sequence. By using a Java based graphical interface the user can visualize the EST maps and the sequence pattern of predicted features.

Homology searches are very important for functional mapping, homology with a known functional region can suggest the function of a query sequence. In particular, when the homologous protein sequence is already known and EST matches are detected, then the gene structure can be reconstructed with high accuracy. Information about EST matches is automatically used by the GeneBuilder system (Milanesi et al., 1999).

Acknowledgment
This work was supported by Italian CNR Genetic Engineering Project

References

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence Identification of 2,375 Human Brain Genes. Nature 355:632--634.
Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McConbie, and J.C. Venter. 1991. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. Science 252:1651--1656.
Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool, J. Mol. Biol., 215, 403-410.
Boguski MS, Tolstoshev CM and Bassett DE, Jr (1994) Gene discovery in dbEST. Science, 265, 1993-1994.
Hillier,L., N. Clark, T. Dubuque, K. Elliston, M. Hawkins, M. Holman, M. Hultman, T. Kucaba, M. Le, G. Lennon, M. Marra, J. Parsons, L. Rifkin, T. Rohlfing, M. Soares, F. Tan, E. Trevaskis, R. Waterston, A. Williamson, P. Wohldmann, and R. Wilson. 1996. Generation and Analysis of 280,000 Human Expressed Sequence Tags. Genome Res., 6, 807-828.
Huang, X (1994) On global sequence alignment. Comput. Applic. Biosci., 10, 227-235.
Jurka J, Walichiewicz J and Milosavljevic A J (1992) Prototypic sequences for human repetitive DNA. J. Mol. Evol., 35, 286-291
Milanesi L., D'Angelo D., Rogozin I.B. (1999) GeneBuilder: interactive in silico prediction of genes
structure. // Bioinformatics, 15, 612-621.
Milanesi L., Rogozin I.B. (1998) Prediction of human gene structure. In: Guide to Human Genome Computing (2nd ed.) (Ed. M.J.Bishop) Academic Press, Cambridge, 215-259.