Gene Recognition Using Multiple Gene-Finding Programs

Katsuhiko Murakami [1] [2] **Toshihisa Takagi** [1]
katsu@ims.u-tokyo.ac.jp takagi@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

[2] Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo 185, Japan

Abstract

This paper demonstrates how effective re-analysis of the results of multiple gene-finding programs is. Four simple algorithms to integrate the results of some programs are proposed and tested. Our experiments show that it is effective to use multiple gene-finding programs simultaneously. We also developed a client program by which one can easily use the algorithm through the Internet.

1 Introduction

A large number of uncharacterized DNA sequences are generated in the development of the genome projects. It is essential to develop algorithms of computational gene finding. Many gene finding programs have been developed. Each program has both good points and bad points in the measures of partial features of genes, such as coding region and splice sites. Therefore, combining the output of several programs may be fruitful if they compensate each other. Moreover, Burset and Guigó discussed that it maybe beneficial to combine the outputs of several gene-finding programs [4]. In this study, we explored the availability of mutual compensation of multiple gene finding programs, such as GRAIL [1], GeneParser [3], GeneFinder [2].

2 Materials and Methods

We extracted human DNA entries with at least one 'CDS' from GenBank Rel. 100 (Apr. 97). We discarded the entries with nonstandard splice sites (not GT-AG), and the entries with the keywords such as pseudo, putative, ORF, alternative, predict, and fusion. The old data registered before June 1996 were discarded. Furthermore, we discarded some sequences whose translated amino acid sequences are very homologous (identity $\geq$ 80 %) with a known protein sequences. The remaining data of 219 loci were used.

After getting prediction by the three programs for the data, we transformed the scores of predicted exons into certain probabilistic scores so that we can compare the quality of the prediction of different programs. We made score functions $Pscore$ defined as $Pscore(score) = A + B \times score$, where $score$ is an output of a gene-finding program, A, B are constants and determined for each program by the error rate distributions.

We considered four different algorithms to combine the results of prediction. In the first method (AND-method), exon candidates are the regions predicted by all the programs. In the second method (OR-method), exon candidates are the regions predicted by any of the programs. In the third algorithm (HIGHEST-method), exon candidates are the region which is given the highest $Pscore$ by one prediction program. The fourth algorithm is based on the performance test shown by Burset and Guigó [4].

In the fourth algorithm (RULE-method), an exon candidate is the same region predicted by the tool selected according to the order of pre-assigned priority. In each method described above, we set threshold to cut exon candidates with low scores. The thresholds were determined so that AC (approximate correlation) was the best. The AC is defined as $AC = \frac{1}{2}\left[\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right] - 1$, where TP is 'True Positive', that is the number of coding nucleotides predicted as coding. FN is 'False Negative', that is the number of coding nucleotides predicted as non-coding. TN is 'True Negative', that is the number of non-coding nucleotides predicted as non-coding. FP is 'False Positive', that is the number of non-coding nucleotides predicted as coding.

3 Results and Discussion

We explored the availability of mutual compensation of multiple gene finding programs, such as FEXH, GeneParser3, and GRAIL. We created four algorithms and tested them. The best AC was 0.76 and it was achieved when all three programs are combined by the method 2 (OR-method), while the performance by single gene-finding program ranged from 0.63 to 0.67. Almost all the methods, except for AND-method, the results of combination of multiple gene-finding programs were better than the results of single program. What is more, there is a tendency that the performance increases as many programs are taken account. With respect to the AND-method, the performance was getting worse as many programs were combined. This results came from the fact that ME became higher (0.45 for FEX+GP+GR) than the extent that WE became lower (0.07 for FEX+GP+GR), where ME (Missing Exons) is the proportion of true exons without overlap to predicted exons and WE (Wong Exons) is the proportion of predicted exons without overlap to actual exons. Despite that we used new DNA sequence data which are not similar to known protein sequences, the combination of the programs made the progress in accuracy. This indicates that the increase of accuracy is not due to just the compensation of individual learned data themselves, but the compensation of feature extraction abilities of different programs.

We have developed a client program 'GeneScope' and a server program 'Shirokane System', to use the methods described here. The details of the GeneScope and the Shirokane System is available via the URL http://gf.genome.ad.jp/.

Acknowledgments

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science" from the Ministry of Education, Science, Sports and Culture, Japan.

References

[1] Y. Xu, J.R. Einstein, R.J. Mural, M. Shah, and E.C. Uberbacher. An improved system for exon recognition and gene modeling in human DNA sequences. In *Intelligent Systems for Molecular Biology*, pages 376–384, 1994.

[2] V.V. Solovyev, A.A. Salamov, and C.B. Lawrence. Identification of human gene structure using linear discriminant functions and dynamic programming. In *Intelligent Systems for Molecular Biology*, pages 367–375, 1995.

[3] E.E. Snyder and G.D. Stormo. Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, 248:1–18, 1995.

[4] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.