

On Gene Prediction by Cross-Species Comparative Sequence Analysis

Rong Chen and Hesham Ali
Department of Computer Science
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE 68182-0116
rchen@mail.unomaha.edu

Abstract

Sequencing of large fragments of genomic DNA makes it possible to perform comparisons of genomic sequences for identification of protein-coding regions. We have conducted a comparative analysis of homologous genomic sequences of organisms with different evolutionary distances and determined the degree of conservation of the non-coding regions between closely related organisms. In contrast, more distance shows much less intron similarity but less conservation on the exon structures. Based on this finding and training of data sets, we proposed a model by which coding sequences could be identified by comparing sequences of multiple species, both close and approximately distant. The reliability of the proposed method is evaluated in terms of sensitivity and specificity, and results are compared to those obtained by other popular gene prediction programs. Provided sequences can be found from other species at appropriate evolutionary distances, this approach could be applied in newly sequenced organisms where no species-dependent statistical models are available.

1. Introduction

Sequencing of large fragments of genomic DNA, and even complete eukaryotic chromosomes, makes it possible to perform comparison of genomic sequences for identification of protein-coding regions. This approach is based on the fact that protein-coding regions evolve much more slowly than non-coding regions. Thus, candidate exons are seen as islands of similarity in alignment of genomic sequences harboring homologous genes.

Recently, several algorithms have been described for automated gene recognition by genomic comparison [1,2,8,9]. Most of these programs are specifically designed for the comparison of the coding regions of closely related species. However, non-coding regions may also be conserved for species whose evolutionary distance is sufficiently close. In our work, we conducted a

comparative analysis of homologous genomic sequences of organisms with different evolutionary distances to find the conservation of the non-coding regions between closely related organisms and to propose a model by which coding sequence could be identified by comparing sequences of multiple species.

2. Comparison of genomic structures with different evolutionary distances

We started by finding out the extent of conservation of genomic structures of different species. For different evolutionary relationships, we chose mouse, chicken, frog and fruit fly for analysis. One hundred and seventeen orthologous human-mouse gene pairs are available from Batzoglu et. al. [2]; TBLASTX was run to search against GenBank to find orthologs in chicken, frog and fruit fly genomic sequences. A dynamic programming algorithm was implemented to determine the similarity among the sequences.

The number and length of exons in human and mouse are well conserved. The number of exons are identical for 95% of total pairs of human and mouse and the lengths are identical in 75%. Between human and chicken, 77% of human and chicken sequences are identical in exon number, and 86% identical in exon length. Compared to the strong conservation of exon-intron structure in human-mouse and human-chicken, for evolutionary distances greater than human-chicken showed less conservation in exon structure.

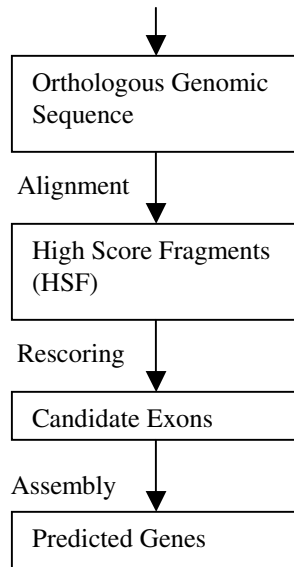
In closely related sequences, all of the aligned coding regions showed strong sequence similarity, non-coding exons and introns also showed conservation in many cases. Therefore, further species could be used together to discriminate coding and non-coding regions.

3. A New Gene Recognition Approach

Base on the results of our comparative research, we sought to develop an automated gene-finding method based on the comparison of multiple organisms with

different evolutionary distances. The model can be presented as follows:

TBLASTX+REPEATMASKER



4. Results and Discussion

We compiled a set of 49 groups of genomic sequences from human, mouse and chicken, which were carefully annotated and used as a standard. Twenty group sequence were used as a training set to optimize the parameters. The remaining human sequences were used for testing purposes. To evaluate the accuracy of our multiple species model, we also tested the same set of data on a model with only human and mouse.

Table 1. Sensitivity and Specificity of our method with human-mouse, human-mouse-chicken comparison, and GeneScan. (Sensitivity = TP/(TP+FN), Specificity = TP/(TP+FP), TP: true positive, correctly predicated; FP: false positive, incorrectly predicated; FN: false negative, missing ones)

	human-mouse model	Human-mouse-chicken model	GeneScan
Sensitivity	0.67	0.83	0.83
Specificity	0.63	0.71	0.75

The result obtained from human-mouse-chicken model shows greater sensitivity and specificity than from human-mouse. Thus, our method demonstrated improved accuracy in gene prediction by applying multiple species comparison of different distances, taking advantage both of the strong conservation of coding region between close species, and divergence of non-coding region between farther species. Our results were comparable to those of

Genescan, the most successful software tool for gene prediction currently available [4,8].

5. Conclusion

Comparative sequence analysis is a powerful approach for detecting functional regions in genomic sequences. Comparing sequences of multiple organisms with both close and approximately distant evolutionarily will increase the reliability of the program. With the increasing number of whole-genome sequencing projects, it will become easy to find syntenic sequence group from organisms with appropriate evolutionary distances. Thus, this method could be applied to predict genomic sequences from newly sequenced organisms and so could be a valuable addition to current gene-prediction tools.

6. References

- [1] Bafna, V. and Huson, D. H. (2000). The Conserved Exon Method for gene finding. In: Proceedings ISMB 8, pp. 3-12.
- [2] Batzoglu, S., Pachter, L., Mesirovi, J. P., Berger, B. and Lander, E. S. (2000). Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* 7, 950-958.
- [3] Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78-94.
- [4] Burset, M. and Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics* 34, 353-367
- [5] Claverie, J.-M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* 6, 1735-1744.
- [6] Guigó, R., Agarwal, P., Abril, J. F., Burset, M. and Fickett, J. W. (2000). An Assessment of Gene Prediction Accuracy in Large DNA Sequences. *Genome Res.* 10, 1631-1642.
- [7] Miller, W. (2001). Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* 17, 391-397.
- [8] Morgenstern, B., Rinner, O., Abdeddaïm, S., Haase, D., Mayer, K., Dress, A. and Mewes, H.-W. (2001). Exon prediction by comparative sequence analysis. In: The Human Genome Meeting 2001, Edinburgh, Programme and Abstract Book pp. 146-147.
- [9] Novichkov, P. S., Gelfand, M. S. and Mironov, A. A. (2001). Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* 17, 1011-1018.