Discovery of protein coding genes through chromosome-to-chromosome sequence comparison

Osamu Gotoh[1,2]
o.gotoh@i.kyoto-u.ac.jp

Masao Morita[3]
m-morita@ybrain.co.jp

Nobuyuki Ichiyoshi[4]
ichiyoshi@mri.co.jp

Tetsushi Yada[1,2]
yt@i.kyoto-u.ac.jp

[1] Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto-shi, Kyoto 606-8501, Japan
[2] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan
[3] Brain Inc., 4-27-11 Kikari, Inzai-shi, Chiba 270-1359, Japan
[4] Information Technology Research Group, Mitsubishi Research Institute, Inc., 2-3-6 Otemachi, Chiyoda-ku, Tokyo 100-8141, Japan

Keywords: gene finding, alignment, comparative genomics, dynamic programming

1 Introduction

Finding genes, especially those encoding proteins, is the first and one of the most important steps toward understanding the information encoded in genomes. In spite of many efforts in the last few decades, computational identification of protein-coding genes in eukaryotic genomes is still in a developing stage. There are three major approaches to this problem: (1) *ab initio* gene finding, (2) methods based on comparative genomic sequences, and (3) methods dependent on known transcriptional products. The comparative approach has been expected to improve both sensitivity and specificity of gene finding compared to the other approaches. However, the heavy computational load has hampered its practical applicability at the chromosomal level. We report here our new program ‘ALNGG’, which makes it feasible to directly compare two full sets of mammalian chromosomes, and automatically identify protein-coding exons on the two genomes. We applied ALNGG to the entire human and mouse genomes. An assessment of the results upon ‘known genes’ on human chromosomes indicated that the sensitivity of our method was comparable to those of the current leading programs [1-3], whereas the specificity was significantly better than any other methods we examined.

2 Methods

Like most previous methods based on cross-species sequence comparison [4, 5], ALNGG adopts a hierarchical strategy. (0) Before applying ALNGG, RepeatMasker [6] is applied to each genomic sequence to filter repeats and low-complexity regions. (1) At the first stage, ALNGG identifies discontinuous oligomers common to the two sequences under comparison. We used a reduced alphabet set to encode conceptually translated genomic sequences. These seeds are extended to both directions to yield ‘initial clumps’. (2) The initial clumps are mutually linked by a sparse dynamic programming (DP) algorithm, which produces a synteny table. (3) Each region within the synteny table with a dense assembly of initial clumps is analyzed by a banded Smith-Waterman-Gotoh algorithm so that it fits to a model of an exon pair of a specific type (initial, internal, last, or single exon). We call such an assembly of initial clumps ‘clumps’. (4) If necessary, a region flanked by adjacent clumps is re-searched under a more relaxed condition of oligomer identity to find weaker clumps. (5) Finally, directions, types and reading frames of a set of clumps are parsed to fit a proper gene model.

All calculations were made on a supercomputer of Human Genome Center at the Institute of Medical Science, The University of Tokyo. It took about seven days on the computer by running 50 CPUs in parallel to finish the analyses with the all chromosomes of human and mouse in both directions.

3 Results and Discussion

We first obtained approximately 10,000 alignments of potentially orthologous pairs of CDS sequences each from human and mouse. By analyzing the matched patterns, we found the most efficient matching pattern of a specific pair of weight (number of matching positions) and width for oligomer identification.

The central procedure of our algorithm is stage 3 in which the boundaries of a homologous pair of exons are identified. The likelihood of an exon pair was measured by the summation of three terms: (1) alignment score assigned to amino-acid substitutions and indels, (2) coding potentials, and (3) boundary signals such as donor and acceptor splicing signals. To calculate the last two terms, we took a simple sum individually obtained from the genomic sequences under comparison.

To assess the performance of ALNGG and other leading programs presently available, we compared the exons predicted by the programs with those annotated in Ensembl Build:34e [7] as 'is_known' (3,913 exons in 461 genes on human chromosome 22). The accuracies of the predictions at the nucleotide sequence level are summarized in Table 1. Although the sensitivities of all the programs were comparable to one another, ALNGG showed significantly better specificity than other programs.

The fraction of exons correctly predicted by ALNGG at the both ends was ~61%., which is significantly better than the results of *ab initio* methods but considerably worse than our transcript-dependent method. In particular, terminal and single exons were difficult to precisely predict. The prediction accuracy will be improved by incorporation of promoter propensities and/or by simultaneous consideration of more than two genomic sequences. Since the performance of a comparative method deeply depends on the distance between the two species and other species-specific characteristics, it is also necessary to develop a procedure to rapidly train the parameter set when ALNGG is applied to a new pair of genomes.

Table 1: Comparison of the performance of representative comparative gene-finding programs.

Program	Sensitivity (%)	Specificity (%)
ALNGG	76	74
Phinal	74	68
SGP	71	60
Twinscan	74	61

References

[1] Noguchi, H., Yada, T., and Sakaki, Y., A novel index which precisely derives protein coding regions from cross-species genome alignments, *Genome Informatics*, 13:183-191, 2003.

[2] Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L., and Dubchak, I. Strategies and tools for whole-genome alignments, *Genome Res.*, 13:73-80, 2003.

[3] Korf, I., Flicek, P., Duan, D., and Brent, M.R., Integrating genomic homology into gene structure prediction, *Bioinformatics*, 17 Suppl 1:S140-148, 2001.

[4] Batzoglou, S., Pachter, L., Mesirov, J.P., Berger B, and Lander, E.S. Human and mouse gene structure: comparative analysis and application to exon prediction, *Genome Res.* 10:950-958, 2000.

[5] Alexandersson, M., Cawley, S., Pachter, L., SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model, *Genome Res*. 13:496-502, 2003.

[6] http://www.repeatmasker.org/

[7] http://www.ensembl.org/Homo_sapiens/index.html