# Coherent Structural Prediction of a Set of Paralogous Genes on a Eukaryotic Genome

**Osamu Gotoh**
gotoh@cancer-c.pref.saitama.jp
Saitama Cancer Center Research Institute
818 Komuro, Ina-machi, Saitama 362, Japan

## 1 Introduction

Following the completion of genomic sequencing of *S. cerevisiae* and *C. elegans*, complete sequencing of several eukaryotic genomes, including that of human, is being accomplished within a few years. An essential but yet unresolved problem is to locate genes on a genomic sequence and to precisely predict their internal (exon-intron) structures. Statistical gene-finding methods have attained significant success, but the performance of even the best available methods is still unsatisfactory for many practical purposes [1, 2]. Homology-based gene-identification methods can considerably improve the accuracy of prediction, provided that one or more known protein or mRNA sequence closely related to the target gene is found in databases [5]. However, it is often observed that the closest relative to a gene is another gene on the same genome. In fact, genomes of higher eukaryotes, such as *C. elegans* and *A. thaliana*, possess a number of large gene families, members of which are mutually well related but far from any genes in other organisms. Here, I propose a method for simultaneously predicting the gene structures of all members in such a species-specific family.

## 2 Methods

The basic idea is similar to that of the "doubly nested randomized iterative strategy" (DNR method) for multiple sequence alignment [4]. We start with a set of translated sequences that have been previously predicted, say by a statistical gene-finding method, and calculate their multiple sequence alignment by the DNR method. Using this alignment as the seed, structures of individual genes are reexamined in turn. Let $A^0_{n,i}(1 \leq n \leq N, 1 \leq i \leq I)$ be the initial multiple alignment, where $N$ is the number of sequences and $I$ is the length of the alignment. To reexamine the gene structure corresponding to the $m$-th sequence, we assign a weight of $w_n = C \cdot pw_{m,n}$ and $w_m = 0$ to sequence $n \neq m$, where $C$ is a normalization factor and $pw_{m,n}$ is the weight for the sequence pair $(m, n)$ in $A^0_{n,i}$ calculated by the three-way method [3]. Optionally, columns predominantly composed of deletion characters may be eliminated. After all of the $N$ sequences are reexamined, a new alignment $A^1_{n,i}$ is constructed from the revised conceptual translation products. This process is repeated until no change in predicted gene organizations is observed.

The program **aln** performs each iterative step based on alignment of a genomic DNA sequence and a protein profile. Another program **prrn** constructs a multiple alignment of translated sequences, $A^k_{n,i}$, in the 'update mode', where the seed for doubly-nested refinement is prepared by substitution of revised sequences for older ones in $A^{k-1}_{n,i}$. The whole process is scheduled by a perl script named **refgs** with essentially no manual intervention.

## 3 Results

Several multigene families in the *C. elegans* genome were examined for the performance of the proposed method. Initial sets of potential family members were taken from those published by The *C. elegans* Sequencing Consortium [6]. When the sequences within a family are moderately related to each other, e.g. the translated products share more than $20 \sim 30\%$ identical amino acids throughout the length, the whole process rapidly converged to yield a multiple alignment with a significantly better sum-of-pairs (SP) or weighted sum-of-pairs (WSP) score than that of the starting alignment. Our method was proven particularly useful for gene-identification of four large families ($> 40$ members) of drug-metabolizing enzymes (cytochrome P450s, UDP-glycosyltransferases, glutathione S-transferases, and short chain alcohol dehydrogenases), since the diversification of these palarogous genes were relatively recent events, and the overall protein architectures are expected to have been well conserved within each family.

## References

[1] Burge, C.B. and Karlin, S., Finding the genes in genomic DNA, *Curr. Opin. Struct. Biol.*, 8:346–354, 1998.

[2] Claverie, J.-M., Computational methods for the identification of genes in vertebrate genomic sequences, *Hum. Mol. Genet.*, 6:1735–1744, 1997.

[3] Gotoh, O., A weighting system and algorithm for aligning many phylogenetically related sequences, *Comput. Applic. Biosci.*, 11:543–551, 1995.

[4] Gotoh, O., Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments, *J. Mol. Biol.*, 264:823–838, 1996.

[5] Gotoh, O., Translated codons (Trons) useful for direct matching of a genomic DNA sequence and a protein sequence or profile, *Genome Informatics 1997*, Universal Academy Press, 266–267, 1997.

[6] The *C. elegans* Sequencing Consortium, Genome sequence of the nematode *C. elegans*: a platform for investigating biology, *Science*, 282:2012–2018, 1998.