The Journal of Microbiology, April 2006, p.137-144
Vol. 44, No. 2
Copyright © 2006, The Microbiological Society of Korea

REVIEW

Computational Approaches to Gene Prediction

Jin Hwan Do[1] and Dong-Kug Choi[2,*]

[1]*Bio-food and Drug Research Center, Konkuk University, Chungju, 380-701, Republic of Korea*
[2]*Department of Biotechnology, Konkuk University, Chungju, 380-701, Republic of Korea*

(Received November 13, 2005 / Accepted November 25, 2005)

The problems associated with gene identification and the prediction of gene structure in DNA sequences have been the focus of increased attention over the past few years with the recent acquisition by large-scale sequencing projects of an immense amount of genome data. A variety of prediction programs have been developed in order to address these problems. This paper presents a review of the computational approaches and gene-finders used commonly for gene prediction in eukaryotic genomes. Two approaches, in general, have been adopted for this purpose: similarity-based and *ab initio* techniques. The information gleaned from these methods is then combined via a variety of algorithms, including Dynamic Programming (DP) or the Hidden Markov Model (HMM), and then used for gene prediction from the genomic sequences.

Keywords: gene prediction, signal/content sensors, similarity-based, gene-finders, *ab initio* gene-finders

The amount of available genome data is increasing exponentially, with the completion of a host of genome sequencing projects. Computational gene recognition programs are critical with regard to the automatic annotation of uncharacterized DNA sequences. Therefore, the perfection of computational techniques for the location of protein-coding regions in uncharacterized genomic DNA sequences constitutes a central issue in the field of bioinformatics. It is, then, crucial to have reliable tools for the automatic annotation of any given DNA sequence of an organism in which the genes and other functions are not currently known, including: the number and location of genes, the location of exons and introns (in eukaryotes), and their exact boundaries. Thus, along with standard molecular methods, a host of new techniques for the determination of distinctive features of protein-coding regions have been proposed over the previous two decades (Mathé *et al.*, 2002). These techniques can be divided into two classes: intrinsic and extrinsic. Intrinsic, or *ab initio,* methods deal strictly with DNA sequences, and extract information regarding gene locations using statistical patterns inside and outside of gene regions, as well as those patterns typical of gene boundaries. Pioneering studies using intrinsic statistical approaches were conducted in the early 1980s (Fickett, 1982; Gribskov *et al.*, 1984; Staden, 1984). Extrinsic or similarity-based methods utilize information derived from similarity search procedures (Robison *et al.*, 1994), using the putative proteins derived from a list of open reading frames (ORFs) as queries. Combinations of both methods have been determined to perform efficiently in terms of gene annotation in large-scale genome sequencing projects (Fleischmann *et al.*, 1995). A host of gene prediction programs are currently available, and the majority of these are referenced in the website maintained by Li (http:// www.nslij-genetics.org/gene).

For several years, we have been conducting a systematic analysis of genomic sequences (human and fungi) via computational prediction and gene cloning (Togashi *et al.*, 2000; Takamatsu *et al.*, 2002; Do *et al.*, 2004; Do *et al.*, 2005a; Do *et al.*, 2005b). Gene discovery in eukaryotic organisms is more difficult than in prokaryotic genomes, due to their low gene density, which is the result of the presence of introns in the coding regions. In this study, we present an overview of approaches to eukaryotic gene discovery,

* To whom correspondence should be addressed.
(Tel) 82-43-840-3610; (Fax) 82-43-840-3872
(E-mail) choidk@kku.ac.kr

and their limitations.

General structure of eukaryotic genes

In eukaryotic protein coding genes, one DNA sequence may code for multiple proteins, due to the presence of alternative promoters or terminators, or of alternative splicing. Splicing patterns are inherently flexible, with variations being observed in different cells and tissues, and at different developmental stages (Maniatis and Tasic, 2002). Alternative splicing significantly complicates *ab initio* computational gene discovery. In general, there no great differences have been observed in the size of protein-coding mRNAs in different organisms, but these genes tend to be larger in vertebrates, and particularly in primates. For example, human coding exons are significantly shorter than the genes. The structural characteristics of eukaryotic genes present several problems with regard to the computational gene identification. The low density of coding regions (3% in human DNA) results in many false positive predictions of non-coding DNA fragments. Small exons (1-20 bp) cannot be recognized using any of the composition-based methods that are successful for prokaryote coding regions. Many *in silico* gene prediction methods have been developed that rely heavily on the recognition of the functional signal encoded within the DNA sequence. These computational methods employ a range of underlying statistical properties of the coding regions, and can be classified as consensus (signal sensors) or non-consensus (content sensors) (Burge and Karlin, 1998; Stormo, 2000). Non-consensus methods can be further classified into trained methods, including GLIMMER (Delcher *et al.*, 1999) and CODONUSAGE (Staden and McLachlan, 1982), which require unbiased sets of coding regions, or untrained methods, including TESTCODE (Fickett, 1982) and GENESCAN (Tiwari *et al.*, 1997), which utilize statistical properties in order to discriminate between coding and non-coding regions.

Signal sensors in transcription, translation and splicing

Many of the important signals that are related to transcription, translation, or slicing have been thoroughly characterized for the prediction of the location and exon/intron organization of genes. The transcriptional signals most often used to locate genes include the initiator or cap signal, located at the transcription start site (TSS), and the A+T-rich TATA box signal, which is normally located at approximately 30 bp upstream of the TSS (Bucher, 1990). Many abundantly expressed genes harbor a strong TATA box in their core promoter. However, in some groups of genes, including housekeeping genes as well as some oncogenes and growth factor genes, no TATA box is present. In the case of translational signals, the "Kozak signal" is often used, primarily due to its lack (in-frame) of coding exons. The detection of transcription and translation signals might facilitate the prediction of the locations of the first and last exons of the genes, but this provides no information regarding the boundaries between exons and introns in a gene. Because most vertebrates, invertebrates, and plant genes tend to harbor several exons, the accurate prediction of gene structure in these organisms relies more profoundly upon the availability of the predictions to pinpoint splice signals (Burge and Karlin, 1998). The dependencies between positions within both the donor and acceptor splice sites have been evaluated via several methods, including 'maximal dependence decomposition' (MDD) and 'windowed weight array' (WWAM) models (Burge and Karlin, 1997). Using a consensus or weighted matrix that reflects the conservative bases of the signals, the high-scoring regions from a given sequence can be selected (Kim and Sim, 2005). Programs for splice site prediction are also available, including SPLICEVIEW (Rogozin and Milanesi, 1997) and SplicePredictor (Kleffe *et al.*, 1996).

Statistical models for content sensing

In general, most current gene prediction programs employ two types of content sensors: one for coding sequences, and one for non-coding sequences such as introns, untranslated regions (UTRs), and intergenic regions. The coding regions possess statistical properties that can help distinguish them from non-coding regions. Several methods have been developed for the identification of putative coding regions within genomic DNA on the basis of nucleotide and codon composition, hexamer frequency, and base periodicity. Fickett and Tung (1992) previously reported that measures predicated on reading-frame-specific hexamer composition resulted in the highest degree of discrimination between coding and non-coding sequences. Therefore, a variety of gene-finding programs, including SORFIND (Hutchinson and Hayden, 1992) and GenView2 (Milanesi *et al.*, 1993), have employed hexamer composition, coupled with a variety of statistical models. The Markov chain model is one of the most frequently utilized statistical approaches. A Markov chain is a sequence of random variables, X_i, in which the probability distribution for X_i relies only on the preceding k variables $X_{i-1},\ldots, X_{i-k}$, for some constant, k. A Markov model of order k captures local dependencies in sequence, at the level of the $k+1$-mers. For example, a fifth-order Markov chain model reflects the dependency in a given hexamer. The larger the order of a Markov model is, the more precisely it is able to characterize dependencies be-

tween adjacent nucleotides. The majority of existing gene prediction programs, including GeneMark (Borodovsky and McIninch, 1993) and GENSCAN (Burge and Karlin, 1997), normally rely on a three-period Markov model of an order of five or less, for the characterization of coding sequences. Salzberg *et al.* (1998) proposed an 'interpolated Markov model' (IMM), which utilizes an interpolation between Markov models of different order. An IMM is able to emulate a fixed *k*th-order chain simply by setting all weights to zero, with the exception of those associated with *k*. IMMs were previously introduced in GLIMMER (Salzberg *et al.*, 1998) and GLIMMERM (Salzberg *et al.*, 1999). The higher-order Markov chain model, or IMM, can be successful in the measurement of nucleotide composition, but these models normally require training using thousands of parameters. This may serve to reduce adaptability, particularly for newly sequenced genomes with small training sets.

Recently, a new system, predicated on Z-curve methods, has been proposed as a method for the recognition of protein-coding genes in bacterial and archaeal genomes (Guo *et al.*, 2003). This method utilizes only 33 parameters and incorporates the coding properties of the protein-coding genes. Thus, it appears to be a suitable method for both eukaryotic and prokaryotic genomes. Below, we briefly summarize the Z-curve method for gene prediction.

Gene prediction via Z-curve

The Z-curve is a unique three-dimensional representation of a DNA sequence, in that the DNA sequence and the Z-curve can each be reconstructed separately and uniquely from the other. The entire Z-curve of a genome allows for both the global and local compositional features to be easily grasped. It is composed of a series of nodes, P_0, P_1, P_2,..., P_N, with the coordinates x_n, y_n, and z_n (in which n = 0, 1, 2, ... , N, and N is the length of the DNA sequence), which are uniquely determined by the Z-transform of the DNA sequence (Zhang and Zhang, 1991; 1994).

$$\begin{aligned} x_n &= (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n \\ y_n &= (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n \\ z_n &= (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n \\ n &= 0, 1, 2, \cdots, N,\ x_n, y_n, z_n \in [-N, N] \end{aligned}$$

Here, A_n, C_n, G_n, and T_n represent the cumulative occurrence positions of A, C, G and T, respectively, in subsequence from the first base to the *n*th base in the sequence. Bases harboring purines and pyrimidines, amino and keto groups, and weak and strong hydrogen bonds are represented by R, Y, M, K, W, and S, respectively. The three components of a Z-curve harbor clear biological import. For example, the component, x_n shows a distribution of the purine/pyrimidine type bases (A or G/C or T) along the sequence. When the number of purine bases in the sub-sequence from the first to the *n*th is greater than that of the pyrimidine bases, $x_n > 0$, otherwise $x_n < 0$. Two groups of samples are required for the recognition of coding regions. One is a set of positive samples corresponding to the true protein coding genes; the other is a set of negative samples corresponding to the intergene sequences. The two groups of samples comprise the training set employed in the Fisher discriminant algorithm. Guo *et al.* (2003) demonstrated that the replication origins and terminations of some bacterial and archaeal genomes could be predicted within the frame of the Z-curve using the x_n, y_n, AT- and GC-disparity curves. Algorithms predicated on the Z-curve have also been employed in the recognition of protein-coding regions within the genomes of eukaryotes (Zhang and Wang, 2000).

Gene prediction via correlation analysis

Although discriminative statistical characteristics derived from training sets tend to be quite successful with regard to the identification of genes from genomic sequences, some genes possess uncharacteristic features, which render this a fairly difficult proposition. Several previous attempts have been made to develop methods for the prediction of coding regions that do require no prior information. A great many of these techniques are predicated on the analysis of correlations within DNA sequences: that is, analysis of the probabilities of locating nucleotides that are separated by a given distance. In general, non-coding sequences appear to have long-range correlations, whereas coding sequences tend to exhibit striking short-range correlations. This phenomenon can be exploited in the development of techniques for the detection of probable genes.

It has been well established that base sequences within the protein-coding regions of DNA molecules possess a period-3 component as the result of the codon structure involved in the translation of base sequences into amino acids (Trifonov and Sussman, 1980). Discrete Fourier Transformation (DFT) is an appropriate technique for the processing of such periodicity. Prior to the application of DFT to a given DNA sequence, numerical values must be assigned to each character (A, T, C, and G) as the DNA sequence is a character string. Voss (1992) assigned a binary sequence to each of the four bases: this allocates a value of 1 to index position n if the corresponding nucleotide is present at that position, and a value of 0 otherwise. The application of DFT to each of these

sequences results in the generation of four spectral representations, which are denoted as $U_A(k)$, $U_T(k)$, $U_C(k)$, and $U_G(k)$, respectively. For a given base b (where b = A, T, C or G), the DFT of the binary sequence $u_b(n)$ of length N is as follows:

$$U_b(k) = \sum_{n=0}^{N-1} u_b(n) e^{-i\frac{2\pi}{N}nk}$$
$$0 \leq k \leq N-1$$

The total frequency spectrum of the given DNA character string is then defined as follows:

$$S(k) = |U_A(k)|^2 + |U_T(k)|^2 + |U_C(k)|^2 + |U_G(k)|^2$$

(Tiwari *et al.*, 1997).

In coding regions of the DNA, the total frequency spectrum $S(k)$ typically exhibits a peak at frequency $k = N/3$, whereas in the non-coding regions, no significant peaks are detected (Chechetkin and Turygin, 1995). The DFT at any particular frequency

Table 1. Gene prediction programs

Program (Website)	Organism	Algorithm*	Alignment	Homology
GeneID (http://www1.imim.es/geneid.html)	Vertebrates, plants	DP, MM		EST
GenLang (http://www.cbil.upenn.edu/~sdong/genlang_home.html)	Vertebrates, *Drosophila*, dicotyledonous plants	Grammar rule		
HMMgene (http://www.cbs.dtu.dk/services/HMMgene)	Vertebrates, *C. elegans*	CHMM		
GLIMMERM (http://www.tigr.org/tdb/glimmerm/glmr_form.html)	Small eukaryotes, *Arabidopsis*, rice	DP, IMM		
GRAIL (http://compbio.ornl.gov/Grail-1.3)	Human, mouse, *Arabidopsis*, rice	DP, NN		
GENSCAN (http://genes.mit.edu/GENSCAN.html)	Vertebrates, *Arabidopsis*, maize	GHMM		
VEIL (http://www.tigr.org/~salzberg/veil.html)	Vertebrates	DP, HMM		
GENIE (http://www.fruitfly.org/seq_tools/genie.html)	*Drosophila*, human, other	GHMM		protein
GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/euk hmm.cgi)	Human, mouse, *Drosophila*, other	GHMM		
FGENESH (http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfind)	Human, mouse, *Drosophila*, rice	HMM		
AAT (http://genome.cs.mtu.edu/aat.html)	Primates, rodents, other		BLASTX/BLASTN	cDNA/protein
GeneSeqer (http://www.maizegdb.org/geneseqer.php)	*Arabidopsis*, maize, generic plant		Spliced alignment	EST/protein
SIM4 (http://pbil.univ-lyon1.fr/sim4.php)	All eukaryotes		BLAST	cDNA/DNA
GeneWise (http://www.ebi.ac.uk/Wise2/index.html)	Human	DP	Global alignment	protein
SLAM (http://baboon.math.berkeley.edu/~syntenic/slam.html)	Human, mouse	DP	Generalized pair HMM	DNA
TWINSCAN (http://genes.cs.wustl.edu/)	Mouse, human	GHMM	BLASTX/BLASTN	DNA

* DP, dynamic programming; MM, Markov Model; NN, Neural Network; HMM, hidden Markov Model; CHMM, class HMM; GHMM, generalized HMM.

reveals a bell-shaped curve surrounding a central value in the coding regions, whereas in the non-coding regions, the distribution is nearly uniform. This regularity has been used to discriminate between coding and non-coding regions within a given non-annotated genomic sequence (Kotlar and Lavner, 2003). This Fourier transform-based technique requires no prior information regarding the genomic structure of the organism, but it has difficulties with regard to the detection of the boundaries of exons and in high noise in small coding regions.

Gene-finders for eukaryotic sequences

Gene-finders have two relevant aspects: one involves the type of information utilized by the program, and the other is the algorithm employed in the combination of that information into a coherent prediction (Stormo, 2000). Depending on the type of information employed, gene-finders can be separated into two classes. Empirical gene-finders, which are also referred to as "sequence similarity-based gene-finders", detect genes via the alignment of known cDNA and protein sequences onto uncharacterized genomic sequences; by way of contrast, *ab initio* gene-finders do not employ sequence similarity, and instead rely on intrinsic gene measures, including signal and content sensors (Yada and Takagi, 2003). In addition to signal and content sensors, similarities can also be used as information. The accuracy of gene prediction programs that use similarity information has shown significant improvements with increases in the numbers of known coding sequences.

Sequence similarity-based gene-finders

The underlying principle inherent to the majority of sequence similarity-based gene-finders is the combination of similarity information with signal sensors. Similarity information can be acquired via a variety of sequence comparisons: genomic DNA/protein, genomic DNA/cDNA, or genomic DNA/genomic DNA. These programs align the genomic DNA sequence against a cDNA database, such as AAT (Huang *et al.*, 1997), GeneSeqer (Usuka *et al.*, 2000), or SIM4 (Florea *et al.*, 1998) (Table 1). This method has proven quite reliable with regard to the identification of exons independently of their coding status, particularly in case the genomic sequence is aligned against a cDNA from the same organism, or a closely related organism (Fukunishi *et al.*, 1999). The comparison of two homologous genomic sequences (inter- or intraspecies) also facilitates the identification of conserved exons, and allows for the simultaneous prediction of genes on both sequences. Programs including SLAM (Alexandersson *et al.*, 2003) and TWINSCAN (Flicek *et al.*, 2003) have also been developed, which exploit the sequence conservation between two genomes in order to predict genes. SLAM utilizes a joint probability model for sequence alignment and gene structure to express the different types of expected alignments, for example, in coding regions and introns. In TWINSCAN, alignments are initially conducted using standard tools such as TBLASTX or BLASTN, and then these alignments are used to inform the prediction algorithms.

Ab initio gene-finders

The majority of *ab initio* gene-finders are predicated on a range of underlying statistical properties of the coding regions, and use a variety of different mathematical techniques, including neural networks, Markov models, and Fourier transforms. They rely on two sequence information types: signal sensors (consensus) and content sensors (non-consensus). A variety of algorithms can be applied to the modeling of gene structure, including Dynamic Programming (DP), Linear Discriminant Analysis (LDA), the Linguist method, Hidden Markov Models (HMM), and Neural Networks. Dynamic programming has often been added into gene-finders, in an effort to combine useful features and facilitate the determination of optimum predictions on the basis of internal scoring systems (Stormo, 2000). Table 1 shows gene-finders and their algorithms, which are used in gene prediction. HMM-based gene-finders have, thus far, proven the most successful in this regard (Guigó *et al.*, 2000). HMMs represent a DNA sequence as the output of an abstract process that progresses through a series of discrete states, some of which are "hidden" from the observer. These states in the context of gene prediction correspond to exons, introns, and any other classes of desired sequences (including 5′ and 3′ UTRs, promoter regions, intergenic regions, and repetitive DNA). The "hidden" aspect of HMMs dictates that we see only the DNA sequence directly, whereas the state that generated the sequence (such as an exon or intron) remains invisible. The output of a regular HMM exhibits a length of 1 for each state within the hidden state space. By this means, a Generalized Hidden Markov Model (GHMM) was developed, in which subsequent states are generated in accordance with a Markov chain, but exhibit arbitrary length distributions. Gene-finders including GENSCAN, GENIE (Reese *et al.*, 2000), HMMgene (Krogh, 2000), and Phat (Cawley *et al.*, 2001) model genomic sequences via a GHMM approach. In general, they determine the sequence states and durations that maximize the joint probability of the hidden and observed data. One of the more attractive features of a GHMM is that it provides an intuitively "natural" method for the computation of the probability

of a predicted exon, given the observed data.

Combination of gene-finders with comparative approaches for gene prediction

Similarity-based gene-finders are able to detect only a limited number of genes (low sensitivity) due to the lack of known mRNAs, whereas *ab initio* gene-finders do not employ sequence similarity, and instead rely on intrinsic gene measures, including coding potentials and splice signals. Two different types of gene-finders, or two or more similar types of gene-finders, can be combined. DIGIT (Yada *et al.*, 2003), for example, generates all possible exons from the results of other gene-finders, such as FGENESH (Salamov and Solovyev, 2000), GENSCAN, and HMMgene, and then assigns them their respective exon types, reading frames, and exon scores; finally, it searches a set of exons whose additive scores are maximized under their reading frame constraints. Another example is EuGène (Schiex *et al.*, 2001): this method uses NetGene2 (Tolstrup *et al.*, 1997) and SplicePredictor for splice site prediction, NetStar (Pedersen and Nielsen, 1997) for translation initiation prediction, IMM-based content sensors, and similarity information gleaned from protein, EST, and cDNA matches. The tracking of exons shared in common by two or more gene-finders carries an advantage in that it significantly reduces the number of over-predictions, but may also exhibit poor sensitivity and possible inconsistencies at the gene level. Platforms such as Genotator (Harris, 1997), MagPie (Gaasterland and Sensen, 1996), and Ensembl (Hubbard *et al.*, 2002) gather evidence acquired from *ab initio* or homology-based prediction programs, and are considered to be relatively useful tools, which facilitate both human-driven and automated annotations.

With such a large number of genome sequencing projects currently underway, the comparative approach is beginning to be seen as more promising in the field of gene identification (Juvvadi *et al.*, 2005). The increasing availability of complete genome sequences makes it possible to conduct a comparison of all of the proteins encoded by one genome with those of another. We have identified the genes relevant to the sphingolipid pathway of *Aspergillus fumigatus*, the genome sequence of which was recently sequenced, via comparative analyses with four other fungi (Do *et al.*, 2005). The predicted genes of *A. fumigatus*

Table 2. The genes predicted to be involved in sphingolipid metabolism in the genome of *A. fumigatus*, and the "best hit" results using BLASTp against the *Saccharomyces* proteome and InterProScan (adapted from Do *et al.*, 2005)

Gene	Contig number/size of predicted protein	Blast results for the yeast proteome: E-value/percent identity; percent similarity in amino acid (aa) overlap	InterPro Accession Number	Molecular function
AUR1	4897 (278,684-279,760)/358 aa	9.7e-70/52%; 70% in 232 aa	IPR000326	PA-phosphatase related
SUR1	4899 (192,183-193,503)/319 aa	4.0e-80/59%; 77% in 233 aa	IPR007577	Glycosyltransferase
CSG2	-	-	-	-
IPT1	-	-	-	-
FEN1	4951 (248,013-249,156)/284 aa	1.4e-31/42%; 57% in 290 aa	IPR002076	Elongation of fatty acids
SUR4	4846 (152,236-150,896)/315 aa	1E-8/24%; 40% in 315 aa	IPR002076	Elongation of fatty acids
TSC13	4840 (47,923-48,592)/223 aa	9.0E-23/35%; 52% in 186 aa	IPR001104	Reductase
LCB1	4941(49,885-51,593)/481 aa	2.6e-68/35%; 53% in 517 aa	IPR004839	Aminotransferase
LCB2	4910 (29,293-31,335)/648 aa	7.4e-155/57%; 72% in 500 aa	IPR004839	Aminotransferase
LCB3	4905 (152,978-154,329)/430 aa	4.1e-19/27%; 47% in 259 aa	IPR000326	PA-phosphatase related
LCB4	4944 (24,440-26,200)/441 aa	2.2e-61/40%; 58% in 316 aa	IPR001206	Diacylglycerol kinase
TSC10	4837 (37,552-38,733)/338 aa	2.7e-10/31%; 42% in 196 aa	IPR002198	Short-chain dehydrogenase/reductase
SUR2	4846 (197,717-199,056)/426 aa	3.8e-80/52%; 68% in 299 aa	IPR006087	SUR2-type hydroxylase/desaturase
DPL1	4840 (101,259-103,208)/559 aa	1.3e-127/48%; 66% in 526 aa	IPR002129	Pyridoxal-dependent decarboxylase
LAG1	4944(103,587-105,786)/400 aa	1e-76/47%; 62% in 327 aa	IPR005547	Longevity-assurance protein
SCS7	4942 (93,723-95,855)/562 aa	5.7e-104/52%; 67% in 385 aa	IPR006694	Fatty acid hydroxylase

exhibited a high degree of similarity to the amino acid sequences encoded by the genes of *Saccharomyces cerevisiae,* and well-conserved functional domains (Table 2). All of the comparative genomic methods harbor a theoretical advantage, in that they are not species-specific. In practice, the performance of these methods depends heavily on the evolutionary distance between the compared sequences. In order to retrieve only the relevant information from homology searches against databases, the use of such programs must be combined with other specific programs that can eliminate repeated sequences, which occur with a fair degree of frequency within the human sequence (about one-quarter of the genome).

Conclusions

Despite the extensive research conducted thus far in the field of gene prediction, current gene prediction programs have not provided a complete solution to the problem of gene identification. For example, short exons tend to be difficult to locate, as discriminative statistical characteristics are less likely to appear in short strands. In addition, the problem of alternative splicing, an important regulatory mechanism in higher eukaryotes, has yet to be effectively resolved. The prediction of protein-encoding genes obviously needs improvement, especially in larger genomes. In order to compensate for the insufficiency of any individual gene prediction program, a computational method for the construction of gene models on the basis of multiple lines of evidence is becoming more feasible. For non-annotated genomic sequences, a diverse set of sources can be combined for annotation, including the locations of gene predictions from *ab initio* gene finders, protein sequence alignments, ESTs and cDNA alignments, and promoter predictions. Such an integrative approach has been demonstrated to consistently outperform even the best individual gene finder and, in some cases, it can result in dramatic improvements in both sensitivity and specificity (Allen *et al.*, 2004).

The computational approach to gene prediction has proven quite useful in terms of gene discovery and related knowledge mining, but biological expertise remains a prerequisite for the confirmation of the existence of a virtual protein, as well as the location or verification of its biological function and the conditions under which it is expressed in the organism.

Acknowledgment

This work was supported in 2004 by the faculty research fund of Konkuk University.

References

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

Allen, J.E., M. Pertea, and S.L. Salzberg. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res.* 14, 142-148.

Borodovsky, M. and J. McIninch. 1993. GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.* 17, 123-133.

Bucher P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 212, 563-578.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78-94.

Burge, C.B. and S. Karlin. 1998. Finding the genes in genomic DN. *Curr. Opin. Struct. Biol.* 8, 346-354.

Cawley, S.E., A.I. Wirth, and T.P. Speed. 2001. Phat–a gene finding program for *Plasmodium falciparum. Mol. Biochem. Parasitol.* 118, 167-174.

Chechetkin, V.R. and A.Y. Turygin. 1995. Size-dependence of three-periodicity and long-range correlations in DNA sequences. *Phys. Lett. A.* 199, 75-80.

Do, J.H., M.J. Anderson, D.W. Denning, and E. Bornberg-Bauer. 2004. Inference of *Aspergillus fumigatus* pathways by comparative genome analysis: tricarboxylic acid cycle (TCA). *J. Microbiol. Biotechnol.* 14, 74-80.

Do, J.H., T.K. Park, and D.-K. Choi. 2005a. A computational approach to the inference of sphingolipid pathways from the genome of *Aspergillus fumigatus. Curr. Genet.* 48, 134-141.

Do, J.H., B.Y. Lim, W.S. Choi, and D.-K. Choi. 2005b. Exploring the Phospholipid Biosynthetic Pathways of *Aspergillus fumigatus* by Computational Genome Analysis. *Eng. Life Sci.* 5(6). 574-579.

Kim, K.B. and J.S. Sim. 2005. Computational detection of prokaryotic core promoters in genomic sequences. *J. Microbiol.* 43, 411-416.

Fickett, J. 1982. Recognition of protein-coding regions in DNA sequences. *Nucleic Acids Res.* 10, 5303-5318.

Fickett, J.W. and C.S. Tung. 1992. Assessment of protein coding measures. *Nucleic Acids Res.* 20, 6441-6450.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science* 269, 496-512.

Flicek, P., E. Keibler, P. Hu, I. Korf, and M.R. Brent. 2003. Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res.* 13, 46-54.

Florea, L., G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8, 967-974.

Fukunishi, Y., H. Suzuki, M. Yoshino, H. Konno, and Y. Hayashizaki. 1999. Prediction of human cDNA from its homologous mouse full-length cDNA and human shotgun database. *FEBS Lett.* 464, 129-132.

Gaasterland, T. and C.W. Sensen. 1996. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* 78, 302-310.

Gribskov, M., J. Devereux, and R.R. Burgess. 1984. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* 12, 539-549.

Guigó, R., P. Agarwal, J.F. Abril, M. Burset, and J.W. Fickett. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* 10, 1631-1642.

Guo, F.B., H.Y. Ou, and C.T. Zhang. 2003. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archael genomes. *Nucleic Acids Res.* 31, 1780-1789.

Harris, N.L. 1997. Genotator: a workbench for sequence annotation. *Genome Res.* 7, 754-762.

Huang, X., M.D. Adams, H. Zhou, and A.R. Kerlavage. 1997. A tool for analyzing and annotating genomic sequences. *Genomics* 46, 37-45.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, *et al.* 2002. The Ensembl genome database project. *Nucleic Acids Res.* 30, 38-41.

Hutchinson, G.B. and M.R. Hayden. 1992. The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res.* 20, 3453-3462.

Juvvadi, P.R., Y. Seshime, and K. Kitamoto. 2005. Genomics reveals traces of fungal phenylpropanoid-flavonoid metabolic pathway in the filamentous fungus *Aspergillus oryzae*. *J. Microbiol.* 43(6). 475-486.

Kleffe, J., K. Hermann, W. Vahrson, B. Wittig, and V. Brendel. 1996. Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. *Nucleic Acids Res.* 24, 4709-4718.

Kotlar, D. and Y. Lavner. 2003. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Res.* 13, 1930-1937.

Krogh, A. 2000. Using database matches with HMMgene for automated gene detection in *Drosophila*. *Genome Res.* 10, 523-528.

Maniatis, T. and B. Tasic. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418. 236-243.

Mathé, C., M-F. Sagot, T. Schiex, and P. Rouzé. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30, 4103-4117.

Pedersen, A.G. and H. Nielsen. 1997. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis, p. 226-233. *In* T. Gaasterland *et al.* (eds). *The Fifth International Conference on Intelligence Systems for Molecular Biology.* AAAI Press, Menlo Park, CA.

Reese, M.G., D. Kulp, H. Tammana, and D. Haussler. 2000. Genie–gene finding in *Drosophila melanogaster*. *Genome Res.* 10, 529-538.

Robison, K., W. Gilbert, and G. Church. 1994. Large-scale bacterial gene discovery by similarity search. *Nat. Genet.* 7, 205-214.

Rogozin, I.B. and L. Milanesi. 1997. Analysis of donor splice signals in different organisms. *J. Mol. Evol.* 45, 50-59.

Salamov, A.A. and V.V. Solovyev. 2000. *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res.* 10, 391-393.

Salzberg, S., A. Delcher, S. Kasif, and O. White. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26, 544-548.

Salzberg, S.L., M. Pertea, A.L. Delcher, M.J. Gardner, and H. Tettelin. 1999. Interpolated Markov models for eukaryotic gene finding. *Genomics* 59, 24-31.

Schiex, T., A. Moisan, and P. Rouzé. 2001. EuGène: an eukaryotic gene finder that combines several sources of evidence, p. 111-125. *In* O. Gascuel and M.-F. Sagot (eds). Lecture Notes in Computer Science, Vol. 2006, First International Conference on Biology, Informatics, and Mathematics, JOBIM 2000. Springer-Verlag, Germany.

Staden, R. 1984. Measurements of the effect that coding for a protein has on DNA sequence and their use for finding genes. *Nucleic Acids Res.* 12, 551-567.

Staden, R. and A.D. McLachlan. 1982. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.* 10, 141-156.

Stormo, G.D. 2000. Gene-finding approaches for eukaryotes. *Genome Res.* 10, 394-397.

Takamatsu, K., K. Maekawa, T. Togashi, D.K. Choi, Y. Suzuki, T.D. Taylor *et al.* 2002. Identification of two novel primate-specific genes in DSCR. *DNA Res.* 9, 89-97.

Togashi, T., D.K. Choi, T.D. Taylor, Y. Suzuki, S. Sugano, M. Hattor *et al.* 2000. A novel gene, DSCR5, from the distal Down syndrome critical region on chromosome 21q22.2. *DNA Res.* 7, 207-212.

Tiwari, S., S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.* 113, 263-270.

Tolstrup, N., P. Rouzé, and S. Brunak. 1997. A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Res.* 25, 3159-3163.

Trifonov, E.N. and J.L. Sussman. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. U.S.A.* 77, 3816-3820.

Usuka, J., W. Zhu, and V. Brendel. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* 16, 203-211.

Voss, R. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* 68, 3805-3808.

Yada, T., T. Takagi, Y. Totoki, and Y. Sakaki. 2003. DIGIT: a novel gene finding program by combing gene-finders. *Pac. Symp. Biocomput.* 375-387.

Zhang, C.T. and J. Wang. 2000. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res.* 28, 2804-2814.

Zhang, C.T. and R. Zhang. 1991. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.* 19, 6313-6317.

Zhang, R. and C.T. Zhang. 1994. Zcurves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.* 11, 767-782.