The Gene-Finder computer tools for analysis of human and model organisms genome sequences

Victor Solovyev[1] and Asaf Salamov

Department of Cell Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030 email:solovyev@cmb.bcm.tmc.edu

Abstract

*We present a complex of new programs for promoter, 3'-processing, splice sites, coding exons and gene structure identification in genomic DNA of several model species. The human gene structure prediction program **FGENEH**, exon prediction - **FEXH** and splice site prediction - **HSPL** have been modified for sequence analysis of Drosophila (**FGENED**, **FEXD** and **DSPL**), C.elegance (**FGENEN**, **FEXN** and **NSPL**), Yeast (**FEXY** and **YSPL**) and Plant (**FGENEA**, **FEXA** and **ASPL**) genomic sequences. We recomputed all frequency and discriminant function parameters for these organisms and adjusted organism specific minimal intron lengths. An accuracy of coding region prediction for these programs is similar with the observed accuracy of **FEXH** and **FGENEH**. We have developed **FEXHB** and **FGENEHB** programs combining pattern recognition features and information about similarity of predicted exons with known sequences in protein databases. These programs have approximately 10% higher average accuracy of coding region recognition. Two new programs for human promoter site prediction (**TSSG** and **TSSW**) have been developed which use Gosh (1993) and Wingender (1994) data bases of functional motifs, respectively. **POLYAH** program was designed for prediction of 3'-processing regions in human genes and **CDSB** program was developed for bacterial gene prediction. We have developed a new approach to predict multiple genes based on double dynamic programming, that is very important for analysis of long genomic DNA fragments generated by genome sequencing projects. Analysis of uncharacterized sequences based on our methods is available through the University of Houston, Weizmann Institute of Science email servers and several Web pages at Baylor College of Medicine.*

Introduction

Large scale genome sequencing projects generate a huge amount of unannotated sequences. Their utility critically depends on the availability of computer methods allowing researchers to extract the maximal possible information from them in a timely fashion. Although the currently existing database searching computer methods allow to infer function based on similarity with known gene or protein, about 50% of the discovered new genes have no detectable homologs in protein databases. Also, among the thousands of expressed sequence tags the number of complete cDNAs is much smaller. This means that pattern recognition methods of gene structure prediction and elucidation of protein function will play a crucial role in annotation of new sequences (Uberbacher et al.,1996).

The recognition of RNA splice sites by the spliceosome is very precise (Aebi,Weismann, 1987; Green,1991) indicating the presence of specific signals for their function. There were many efforts to analyze the sequences around conserved regions of donor and acceptor splice sites (Breathnach, Chambon,1981; Mount,1982; Senapathy et al.,1990). Scoring schemes based on consensuses or weight matrices, free energy of base-pairing of RNA with snRNA and other peculiarities, give an accuracy of about 80% for the predicting splice site positions (Nakata et al.,1985; Gelfand,1989). More accurate prediction (95%) is shown by neural network algorithms (Lapedes et al.,1988; Brunak et al.,1991). In our investigation a Bayesian prediction scheme for detection of splice sites has been used. We designed linear discriminant functions (combining seven different characteristics) for human donor and

1. Present address: Amgen Inc. MS 14-1-D 1840 DeHavilland Drive,Thousand Oaks, CA 91320-1789

acceptor splice site identification and achieved 96-97% accuracy (Solovyev, Lawrence, 1993; Solovyev et al., 1994).

Most gene prediction systems combine information about functional signals and the regularities of coding and intron regions. The program ***SORFIND*** (Hutchinson, Hayden, 1992) was designed to predict internal exons based on codon usage and Berg & von Hippel (1987) discrimination energy for intron-exon boundaries recognition. An accuracy of exact internal exons prediction (at both 5' and 3' splice junctions and in the correct reading frame) by SORFIND program reaches 59% with a specificity of 20%. A dynamic programming approach (alternative to the rule-based approach) was applied by Snyder and Stormo (1993) to internal exon prediction in GeneParser algorithm. GeneParser recognized 76% of internal exons, but the structure of only 46% exons was exactly predicted when tested on entire GenBank entry sequences. Recently, the Genie system using a Generalized Hidden Markov Model has been developed. It is similar in design to GeneParser, but is based on a rigorous probabilistic framework (Kulp et al.,1996). We have developed a program (***HEXON***) based on a splice site prediction algorithm and preferences of oligonucleotides in protein coding and intron regions (Solovyev et al., 1994). ***HEXON*** had a better exact exon prediction quality than the other internal exon prediction programs and may be useful for analysis of partially sequenced genes.

A number of gene structure prediction programs have been developed to assemble potential eukaryotic coding regions into translatable mRNA sequence selecting optimal combinations of compatible exons (Fields & Soderlund,1990; Gelfand,1990; Guigo et al.,1992; Dong & Searls,1994). Dynamic programming was suggested as a fast method to find an optimal combination of preselected exons (Gelfand, Roytberg, 1993; Solovyev, Lawrence, 1993b; Xu et al.,1994), that is different from the approach suggested by Snyder and Stormo (1993) to search for exon-intron boundary positions. We have combined 5'-, internal and 3'-exon identification linear discriminant functions and dynamic programming approaches in our gene prediction system (Solovyev et al.,1994-95). A comprehensive test of our and the other gene finding algorithms has been made recently by the developers of ***GeneID*** algorithm (Burset, Guigo,1996). Our GeneFinder program (***FGENEH***) (Solovyev et al.,1995) is one of the best in the tested group having the exact exon prediction accuracy 10% higher than for the others and the best level of accuracy on the protein level. The same best performance of ***FGENEH*** was shown by the developers of ***Genie*** (Kulp et al.,1996). The above mentioned tests were done without using any information about possible similarity of analyzed sequence and a known sequence in the data base. If the similarity exist, ***GeneID***, ***GeneParser*** and ***GeneFinder*** variants can exploit this information and they show much better accuracy in this case.

While significant success has been made in coding region identification, however, perfect prediction of eukaryotic gene structure continues to be a challenging problem to the biologist trying to deduce the translated proteins from the genomic sequences. Also, we need to develop specific programs for analysis of genomic DNA from different species as long as it has become clear that the rules of splicing might be dissimilar in different taxonomic classes (Mount, 1993). There are several genome sequencing projects have been launched for different organisms, therefore it is extremely important to develop software for analysis of their sequences. This work is mostly devoted to extension of our human GeneFinder tools to identification of splice sites, exons and gene-structure in several model organisms as well as prediction of multiple genes in long genomic DNA.

Results and Discussion.

Predicting splice sites in model organisms

We have used the same characteristics for splice site identification which were selected during our analysis of human genes (Solovyev, Lawrence, 1993; Solovyev et al.,1994). The characteristics computed for classifying donor site are: the triplet preferences in the potential coding region (-30 to -5), conserved consensus region (-4 to +6) and G-rich region (+7 to +50); the number of significant triplets in the conserved consensus region; octanucleotide preferences of being coding in the -60 to -1 region and intron in the +1 to +54 region; and the number of G-bases, GG-doublets and GGG-triplets in the +6 to +50 region. The characteristics for acceptor splice sites are: the triplet preferences in the branch point region (-48 to -34), poly(T/C)-tract region (-33 to -7), conserved consensus region (-6 to +5), coding region (+6

to + 30); octanucleotide preferences of being coding in the (+1 to +54) region and intron in the (-1 to -54) region; and the number of T and C in poly(T/C)-tract region. We learned statistical parameters of our splice site discriminant functions from Drosophila, Nematode, Yeast and Plant (Arabidopsis) genomic sequences. The accuracy of splice site recognition by ***DSPL, NSPL, YSPL and ASPL*** programs is estimated on the genes that have not been included in the learning set (Table 1). A group of gene sequences presented to the GenBank in 1996 were taken as a test set (here and in the following sections), while genes deposited before that year were used as a learning set.

Table 1. The accuracy of splice sites identification.

Organism	splice site	#examples	(Sn + Sp)/2	Corr. coeff.
Drosophila	donor	874/287	95.5%	0.61
Drosophila	acceptor	883/287	95.5%	0.62
Nematode	donor	6761/856	95.0%	0.59
Nematode	acceptor	6770/841	95.0%	0.60
Yeast	donor	845/14	96.0%	0.67
Yeast	acceptor	856/14	90%	0.46
Plant	donor	497/1009	95.5%	0.65
Plant	acceptor	499/988	93.0%	0.56

Shown the number of examples in the learning/test sets; Sn (sensitivity), Sp (specificity) and correlation coefficient are usual measures of accuracy; (Sn+Sp)/2 was suggested as an integral measure in (Fickett,Tung,1993)

We observe enough good accuracy of splice site prediction for analyzed model organisms. It is similar with our results received for human genes. However, we can see small decrease of quality that may hint to search for new features specific for these organisms. A new plant splice site prediction method combining neural network and rule based approaches was designed recently (Hebsgaard et al.,1996). It combines local and global information and its accuracy should be compared with results of gene-structure prediction programs.

Predicting 5'-, internal, 3'- coding exons.

For exon prediction we also took the same characteristics of discriminant functions that were selected for human genes. During internal exon recognition we use the octanucleotide preferences of being intron in the potential left intron region; the value of the acceptor splice site recognition function, the ORF octanucleotide preferences of being coding, the value of the donor splice site recognition function and the octanucleotide preferences of being intron of the right potential intron region.

As 5'-exon candidates we consider all open reading frames that starting with an ATG codon and ending with a GT dinucleotide. As components of the 5'-exon recognition function we take the hexanucleotide preferences for 5'-regions -150 - -101 bp, -100 - -51 bp, -50 - -1 bp of being to the left of the potential coding region; the average value of positional triplet preferences in the -15 - +10 region around ATG codon; octanucleotide preferences to be coding region of ORF, the value of donor splice site recognition function and the octanucleotide preferences of being intron of the right potential intron region.

As potential 3'-exons we consider all ORF regions starting after AG base pair and finishing with a stop codon. As components of the 3'-exon recognition function we take the octanucleotide preferences of being intron of the left potential intron region; the value of the donor splice site recognition function; octanucleotide preferences in ORF to be coding region, hexanucleotide composition preferences for 3'-region +1- +50 bp, +51 - +100 bp, +101 - +151 bp

to the right of the potential coding region; the average value of positional triplet preferences in the -10 - +30 region around the stop codon.

We learned statistical parameters of our exon discriminant functions from known gene sequences of different species. The accuracy of exon recognition by programs ***FEXD, FEXN*** and ***FEXA*** for the test set sequences is presented in Table 2.

Table 2. The accuracy of exon identification.

Organism	exons	#examples	(Sn + Sp)/2	Corr. coeff.
Drosophila	internal	166/78	94.5%	0.90
Drosophila	all	478/152	87.0%	0.80
Nematode	internal	4528/423	95.0%	0.91
Nematode	all	6562/605	92.0%	0.88
Plant	internal	287/443	94.2%	0.89
Plant	all	457/665	93.5%	0.79

We observe a good accuracy of internal exon prediction (94-95%). But flanking exons are predicted much less accurate. This is a common feature of many exon prediction programs. For example, the GeneParser internal exon prediction tested on sequences between the first and the last exons had correlation coefficient (C) about 0.8. However, prediction on the whole gene sequences resulted with C=0.68 (Snyder,Stormo,1993). We suppose that this effect is due to poor identification power of 5'- and 3'-end characteristics in exon discriminant functions. Their improvement is one of the main task of future work in gene recognition.

Prediction of gene structure.

We applied dynamic programming to predict optimal gene model from the list of potential exons. The method have been described for human sequences in (Solovyev et. al.,1995). The same procedure was implemented in programs of gene structure prediction for model species. Only minimal intron length was changed to 30 bases and organism-specific splice site and exon prediction functions described above were used. The results of applying gene identification programs ***FGENED, FGENEN*** and ***FGENEA*** to the test set of sequences are presented in Table3.

Table 3. The accuracy of gene structure prediction.

Organism	# test examples	(Sn + Sp)/2	Corr. coeff.
Drosophila	152	90.5%	0.86
Nematode	605	92.0%	0.88
Plant	665	90.5%	0.86

These programs have usually a better accuracy than the total accuracy for separate exons (5', internal and 3') prediction. However, they can produce a wrong result if a sequence contained many genes. In this case it would be better to search for exons without their assembling or apply new technique considered below.

Using database homology search

We have developed the versions of ***FGENEH*** and ***FEXH*** (referred as ***FGENEHB*** and ***FEXHB***) which combine pattern-recognition exon prediction and information about similarity between translated amino acid sequences of the predicted exons and

known sequences in a protein database. Addition of homology information often may significantly increase an accuracy of gene-finding programs, especially when a close homologs occur in the database (Burset, Guigo, 1996). Recently some variants of gene prediction programs exploiting a similarity search have been developed, such as ***GeneID+*** (Guigo et al, 1992), ***GeneParser+*** (Snyder, Stormo, 1995) and ***GRAIL*** (Xu,Uberbacher, 1996).

In ***FGENEHB*** and ***FEXHB*** we use a lower thresholds for the corresponding discriminant functions than in ***FGENEH*** and ***FEXH***, to allow analyze a greater number of potential exon candidates. Database searches with the translated amino acid sequences of predicted exons is performed by the FASTA program (Pearson, Lipman, 1988). To decrease the search time, we create one amino acid sequence for each cluster of overlapping exons with the same translation frame. In the current implementation of ***FGENEHB*** and ***FEXHB*** we only modified the original weights of predicted exons by adding an additional term which is proportional to the observed homology level. From the FASTA output beginning with the first hit we extract homology level between our query sequence (in this case the translated amino acid sequence of the potential exon) and the high-scoring database sequences selected by FASTA. If some fragments of query sequence is not covered by the first hit, the homology level is extracted from the subsequent 4 best hits. After that we compute the average homology level for each potential exon. The original weights of these exons is augmented by the additional term, non-linearly proportional to the average homology level (higher scaling factors for the higher homology level). Table 4 presents the results of ***FGENEHB***'s performance on a dataset of 570 entries compiled by Burset and Guigo (1996) in comparison with the results of GeneID+ and GeneParser 3 which also used homology information from the protein database.

Table 4. Performance of gene-recognition programs, using homology information from the protein databases for the dataset of 570 GenBank entries (Burset & Guigo, 1996)

	Nucleotide				Exact exon			
	Sn	Sp	Ac	Cc	Sn	Sp	ME	WE
Fgenehb+	0.92	0.96	0.92	0.92	0.74	0.78	0.09	0.02
GeneID+++	0.91	0.91	0.88	0.88	0.73	0.70	0.07	0.13
GeneParser3++	0.86	0.91	0.86	0.85	0.56	0.58	0.14	0.09

+) In the testing the first FASTA hit was not considered, to remove possible exact protein product of the gene. ++) The values for GeneID+ and GeneParser3 are taken from the article of Burset and Guigo (1996).

This simple way of accounting similarity information could be used with EST database also.

Prediction of multiple genes.

Currently there is no program that can predict several genes without using any information about the structure of known similar sequences. The behavior of known gene-finding algorithms is unpredictable in this case. They could either predict genes containing exons from the other genes or they could identify the only one right gene having the highest score.

We applyed double dynamic programming to develope a new gene structure prediction algorithm by analyzing many possible gene models in a given sequence and selecting some optimal combination of them. Let consider a short description of this approach. Each predicted exon is characterized by its weight (the LDF value) and reading frame {0,1,2}. An array of such exons ordered according their start positions could be presented as vertices in the directed acyclic graph. We will call any 2 exons as compatible (i.e. connected by an edge in the graph) if: 1) the first exon is either 5' or internal exon and the second exon is either internal or 3'-exon; 2) the distance between the first exon end and the second exon start is more than the minimum intron length (60bp in human genes); 3) the ORFs of these exons are compatible upon their merging, i.e after removing the corresponding intron sequence the combined ORF will contain no in-frame stop codons. Our algorithm consists of 2 successive dynamic programming steps. During the first step for each exon pair (the first exon

is either 5' or internal and the second one is internal or 3') we look for an optimal gene model beginning with the first exon and ending with the second (optimal model is one with the maximal total weight). This problem corresponds to the well-known problem of finding shortest (or longest) path between all pairs of vertices in the directed acyclic graph (Cormen et al.,1990) and can be solved by dynamic programming. If N is the number of preselected exons, then the maximal number of all considered gene models is (N-1)*(N-2)/2 (if all exons are internal and non overlapping with each other). In practice the actual number of gene models is much lower, because: 1) we analyze only the models with an average exon weight higher than a certain threshold; 2) we exclude all models which are subsets of larger models; and 3) only compatible exons could be connected in the graph. During the second step we consider an array of potential gene models which ordered according their start positions as vertices of new directed acyclic graph. Each such model is characterized by the total weights of exons belonging to it. We define any 2 gene models as compatible if: 1) the distance between them is greater than a minimal distance between 2 consecutive genes observed in human sequences; 2) they are not covered by a larger model with an average exon weight higher than the average weights for the small models. We also use dynamic programming to find the model combination (or the path in the directed acyclic graph) with maximal total weight. The vertices of this path correspond to the potential gene models.

We implemented this double dynamic programming approach in ***FGENEM*** program and tested it on several long sequences containing gene clusters. As the first example the well known HUMBB GenBank sequence of 73308 bases including five β-globin genes and one pseudogene was selected. The ***FGENEM*** program predicted 5 genes in this sequence. g-gamma, a-gamma and beta-globin genes were predicted exactly; epsilon-globin gene was identified almost right, only its 5'-exon was 18 amino acid shorter at the beginning; in delta-globin gene the last exon was predicted exactly, the second exon partially and the first exon was missed. So, we correctly defined the number of potential genes, and exactly predicted positions of 12 exons, 2 exons were predicted partially and only 1 was not found in this long sequence.

The next test example was taken from the recent work of Xu and Uberbacher(1996) where the authors presented their approach to predict multiple genes base on information about known homologous protein sequences in the database. They created artificial sequence by combining three one-gene GenBank entries (HUMCYPIIE of 9 exons; HUMRASH of 4 exons and HUMACTGA of 5 exons). Our program found exactly 3 genes in this sequence. The structure of HUMRASH gene was predicted perfectly. 4 of 5 exons of HUMACTGA were found correctly and one was defined by 2 amino acid shorter than real. Six exons of HUMCYPIIE were predicted exactly, 2 were overlapped and one was missed. Our results is not significantly different from the prediction approach that used information about similar sequences in protein database. That approach illustrated only on this example also missed one exon and predicted one extra (Xu & Uberbacher,1996). We understand that there is a lot of work could be done to provide good empirical parameters for this algorithm, however considered above preliminary results encourage its future development.

Recognition of promoter regions in human DNA.

Eukaryotic polymerase II promoter sequences are the main functional elements of eukaryotic genes. Their recognition by computer algorithms will increase the quality of gene structure identification as well as provide possibility to study gene regulation. However, development of computer algorithms to recognize Pol II promoter sequences in genomic DNA is an extremely difficult problem of computational molecular biology. Promoter 5`-flanking region is very poorly described in general. It may contain dozen short motifs (5-10 bases) that serve as recognition sites for proteins providing initiation of transcription as well as specific regulation of gene expression. These motifs are different in various groups of genes and even such known promoter element as TATA-box is often absent in 5'-regions of many house-keeping genes. Each promoter has a unique selection and arrangement of these elements providing a unique program of gene expression. Currently, the only one known program (***PromoterScan***) can predict promoter regions with a relatively small false positive rate (Prestridge,1995). The program searches for eukaryotic promoter regions using a motif data set and can recognize about 50% of test promoter sequences.

Recently we have developed 2 new programs

(***TSSW*** and ***TSSG***) for eukaryotic promoter identification. They analyze occurrence of functional motifs from 2 databases (Ghosh,1993) and (Wingender,1994), respectively and sequence composition characteristics of potential promoter regions and transcription start sites (TSS). The main compositional characteristics are positional triplet preferences around the potential start of transcription; hexaplet preferences in the regions -1 - -101, -101- -201 and -201 - -301 relative to TSS; the score of TATA box weight matrix. The programs can predict positions of transcription initiation and binding sites of transcription factors for approximately 50% of known promoters with one false positive prediction for about 5000 bases of genomic DNA. They provide a better average accuracy of locating TSS positions than the previously developed method (Prestridge,1995). For example, when we analyzed 10 promoter recognizable by both methods (***TSSG*** and ***PromoterScan***) the average deviation of predictions by ***TSSG*** program from the real TSS was about 10 bases, while for ***PromoterScan*** predictions this characteristic was 70 bases.

Recognition of 3'-regions of human genes.

One of the distinguishing fragments of eukaryotic transcripts is the 3'-untranslated region (3'UTR) which has the diversity of cytoplasmic functions affecting the localization, stability and translation of mRNAs (Decker and Parker, 1995). Recently there have been a few attempts to predict 3'-processing sites by computational methods. Yada et al (1994) have conducted a statistical analysis of human DNA sequences in the vicinity of poly-A signal in order to distinguish them from nonactive in polyadenylation AATAAA sequences occurring in human DNA (pseudo polyA-signals). They found that a base C frequently appears on the upstream side of the AATAAA signal and a base T or C often appears on the downstream side, implying that CAATAAA(T/C) can be regarded as a consensus of the poly-A signal. Kondrakhin et al (1994) constructed a generalized consensus matrix using 63 sequences of cleavage/polyadenylation sites in vertebrate pre-mRNA. The elements of their matrix were absolute frequencies of triplets at each site position. Using this matrix they have provided a multiplicative measure for recognition of polyadenylation regions. However their method has a very high false positive rate.

We have developed a computer program ***POLYAH*** and algorithm for identification of 3'-processing sites of human mRNA precursors. The algorithm is based on a linear discriminant function (LDF) trained to discriminate real poly-A signal regions from the other regions of human genes possessing the AATAAA sequence which is most likely nonfunctional. As the parameters of LDF various significant contextual characteristics of sequences surrounding AATAAA signals were used. We used for analysis a 300 nucleotides sequence region (-100, +200) around AATAAA hexamer (the location of the first base of AATAAA is defined as position 1). Pseudo sites were taken out of human genes as the sequences comprising (-100,+200) around the patterns revealed by poly-A weight matrix (see below), but not assigned to poly-A sites in the Feature table. Sequences submitted to GenBank before 1994 were included to the training set and those after 1994 to the test set. As a result there were 248 positive and 5702 pseudosites in the training set and respectively 131 and 1466 in the test set. As characteristics of poly-A discriminant function we selected: 1) Position weight matrix for scoring of poly-A signal; 2) Position weight matrix for scoring of downstream GT/T-rich element; 3) Distance between poly-A signal and predicted downstream GT/T-rich element; 4) Hexanucleotide composition of downstream (+6,+100) region; 5)Hexanucleotide composition of upstream (-100,-1) region; 6) Positional triplet composition of downstream (+6,+55) region; 7) Positional triplet composition of upstream (-50,-1) region, 8)Positional triplet composition of the GT/T-rich downstream element. Detailed description of these features and the method was presented in (Salamov and Solovyev, 1997).

The most significant characteristic is the score of AATAAA pattern (estimated by the position weight matrix) that indicates the importance of occurrences of almost perfect poly-A signal (AATAAA). The second valuable characteristic is the hexanucleotide preferences of downstream (+6,+100) region. Although the discriminating ability of GT-rich downstream element itself (characteristic 2) is very weak, combining it with the other characteristics significantly increases the total Mahalonobis distance.

When the threshold was set to predict 86% of poly-A regions correctly, specificity of 51% and correlation coefficient of 0.62 had been achieved. The precision of this approach is better than for the other methods (see, for example, a recent one by Kondrakhin et al.,1994) and has been tested on a larger data set.

Kondrakhin et al (1994) did not report the total prediction accuracy of their method. Instead, they tested their recognition function for the sequence of adenovirus (Ad2) genome. This sequence (35937 bp) contains poly-A sites for nine transcripts which are polyadenylated at different stages of Ad2 onthogenesis. Authors reported the error rates of their method at different thresholds for poly-A signal selection. If the threshold is set to predict 8 of 9 real sites their function also predicts 968 additional false sites. We have tested the ***POLYAH*** program with the same sequence of Ad2 genome and for 8 correctly predicted sites it gives only 4 false sites. In general, the accuracy of ***POLYAH*** is significantly better than provided by previous two methods. While about 86% of real poly-A regions can be predicted correctly, the more precise recognition will need an additional study of these complex gene regions.

Prediction of bacterial genes.

A new method has been developed to find protein coding genes in E.coli DNA. The method is based on the discriminant analysis of open reading frames flanked ATG(GTG) and STOP codon pairs. Prediction is performed by linear discriminant function combining characteristics describing 5'-, 3'-mRNA regions and coding region for each open reading frame: octanucleotide preferences for coding region; in-frame octanucleotide preferences for coding region; hexanucleotide prefernces of -30 - +10 region around (ATG/GTG) codons; hexanucleotide prefernces of -10 - + 30 region around stop codons; positional triplet preferences around (ATG/GTG) codons and positional triplet preferences around potential stop codons. The parameters of the linear discriminant function was computed using 3000 E.coli sequences of annotated DNA in GenBank (1994) entries. The performance of bacterial gene prediction program ***CDSB*** was estimated on 1000 genes presented to GenBank in 1995. The recognition quality computed at the level of individual nucleotides is 94% and 75% for precise recognition of gene locations.

The methods availability.

Our computer programs for sequence analysis are installed in several genome sequencing centers such as U.C. Berkeley, Columbia University, Washington University, National Institute of Health, Wisconsin Medical College, UK MRC HGMP Resource Centre and in 2 email servers (University of Houston and Weizmann Institute of Science). We have developed several WWW pages using HTML and Perl scripts. For example, "Lost protein trail" Web page: (http:// ***defrag.bcm.tmc.edu:9503/lpt.html***) unites 21 programs for splice site, promoter and gene structure prediction for different species; 4 programs for protein secondary structure and prosite pattern prediction and programs for fold recognition. This WWW page usually provides new version of programs as comparing with Email servers and others WWW sites (see, for example, http:// ***dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html*** and http:***//dot.imgen.bcm.tmc.edu:9331/psspredic-tion/pssp.html***).

References.

Aebi M., Weissmann C. (1987) Precision and orderliness in splicing.Trends in Genetics, 3, 102-107.

Breathnach R.,Chambon P. (1981) Organization and expression of eukaryotic split genes for coding proteins. Ann.Rev.Biochem., 50,349-393.

Berg, O.G., von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins. J.Mol.Biol., 193, 723-750.

Burset M., Guigo R. (1996), Evaluation of gene structure prediction programs, Genomics, 34(3), 353-367.

Cormen T.H., Leiserson C.E., Rivest R.L. 1990. Introduction to Algorithms. MIT Press.

Decker C.J. and Parker R. (1995) Diversity of cytopasmatic functions for the 3'-untranslated region of eukaryotic transcripts. Current Opinions in Cell Biology 1995, 7: 386-392.

Dong S., Searls D.(1994), Gene structure prediction by linguistic methods, Genomics, 23, 540-551.

Brunak, S.; Engelbreht J.; Knudsen S. (1991) Prediction of Human mRNA donor and acceptor sites from the DNA sequence. J. Mol.Biol. 220: 49-65.

Efron B., Tibshirani R. (1991) Statistical data analysis in the computer age. Science, 253, 390-395.

Fickett J.W.; Tung C.S. 1992. Assessment of Protein Coding Measures. Nucl. Acids Res. 20: 6441-6450.

Fields C.,Soderlund C. (1990) GM: a practical tool for automating DNA sequence analysis", CABIOS, 6, 263-270.

Gelfand M. (1989) Statistical analysis of mammalian pre-mRNA splicing sites, Nucleic Acids Research, 17, 6369-6382.

Gelfand M. (1990), Global methods for the computer prediction of protein-coding regions in nucleotide sequences, Biotechnology Software, 7, 3-11.

Gelfand M.,Roytberg M. (1993), Prediction of the exon-intron structure by a dynamic programming approach,

BioSystems, 30(1-3), 173-182.
Ghosh, D. (1993). Status of the transcription factors database (TFD). Nucl. Acids Res.21, 3117-3118.
Green M.R. (1991) Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. Ann.Rev.Cell Biol.,7,559-599.
Guigo R.; Knudsen S.; Drake N.;Smith T.(1992) Prediction of gene structure.J.Mol.Biol.226:141-157.
Hutchinson G.B., Hayden M.R. 1992. The prediction of exons through an analysis of splicible open reading frames. Nucl.Acids Res. 20:3453-3462.
Hebsgaard S., Korning P., Tolstrup N., Engelbrecht J., Rouze P., Brunak S. (1966) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucl.Acids Res., 24,3439-3452.
Kondrakhin Y, Shamin V., Kolchanov N. (1994) Construction of a generalized consensus matrix for recognition of vertebrate pre-mRNA 3'-terminal processing sites. Comput. Applic. Biosci., 10, 597-603.
Kulp D., Haussler D., Reese M., Eeckman F.(1996) A generalized hidden Markov model for the recognition of human genes in DNA, in In Proceedings of the 4th International Conference on Intelligent Systems for Molecular Biology, 134-142.
Lapedes A.; Barnes C.; Burks C.; Farber R.;Sirotkin K. (1988) Application of neural network and other machine learning algorithms to DNA sequence analysis. In Proceedings Santa Fe Institute 7: 157-182.
Mount S.M. (1982) A catalogue of splice junction sequences. Nucl.Acids Res. 10: 459-472.
Mount S.M. (1993) Messenger RNA splicing signal in Drosophila genes. In An atlas of Drosophila genes (ed. Maroni G.), Oxford.
Nakata K.; Kanehisa M.; DeLisi C. (1985) Prediction of splice junctions in mRNA sequences. Nucl.Acids Res. 13: 5327-5340.
Prestridge D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. J. Mol. Biol. 249: 923-932.
Senapathy P.; Shapiro M.B.; Harris N.L.(1990) Splice junctions, Branch point sites, and Exons. Methods of Enzymology (ed. R.F. Doolittle) 183: 252-280.
Snyder E.E.,Stormo G.D. (1993) Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. Nucl.Acids Res., 21:607-613.
Salamov A.A., Solovyev V.V. (1997) Recognition of 3'-processing sites of human mRNA precursors. CABIOS 13, 23-28.
Solovyev V.V., Lawrence C.B. (1993) Identification of Human gene functional regions based on oligonucleotide composition. In Proceedings of the First International Conference on Intelligent Systems for Molecular Biology (eds. Hunter L.,Searls D.,Shavlik J.), p.371-379.
Solovyev, V., Lawrence, C. (1993b) Prediction of human gene structure using dynamic programming and oligonucleotide composition In: Abstracts of the 4th annual Keck symposium. Pittsburgh, 47.
Solovyev V.V., Salamov A.A., Lawrence C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of splicible open reading frames. Nucleic Acids Res. 22,N 24, 5156-5163.
Solovyev V.V., Salamov A.A., Lawrence C.B. (1995) Prediction of human gene structure using linear discriminant functions and dynamic programming. In Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (eds. Rawling C.,Clark D.,Altman R.,Hunter L.,Lengauer T.,Wodak S.), Cambridge,England, AAAI Press,367-375.
Uberbacher E.C., Xu Y., Mural R.J. (1996) Discovering and understanding genes in Human DNA sequences using GRAIL. in Methods Enzimology (ed. Doolittle R.), 266, 259-281.
Wingender, E. (1994) Recognition of regulatory regions in genomic sequences. J.Biotechnol. 35, 273-280.
Xu Y., J.R. Einstein, R.J. Mural, M. Shah and E.C. Uberbacher (1994) An improved system for exon recognition and gene modeling in human DNA sequences". In Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology (eds. Altman R., Brutlag, Karp P, Lathrop R. and Searls D.) 376-383.
Xu Y., Uberbacher E. (1996) Gene prediction by pattern recognition and homology search. In Proceedings of the 4th International Conference on Intelligent Systems for Molecular Biology, 241-252.
Yada T., Ishikava M., Totoki Y., Okubo K. (1994) Statistical analysis of Human DNA sequences in the vicinity of poly-A signal. In Proceedings of the 2-nd International conference on Bioinformatics, Supercomputing, and Complex genome analysis.