Identification of human gene structure using linear discriminant functions and dynamic programming

Victor V. Solovyev, Asaf A. Salamov and Charles B. Lawrence

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

Abstract

Development of advanced technique to identify gene structure is one of the main challenges of the Human Genome Project. Discriminant analysis was applied to the construction of recognition functions for various components of gene structure. Linear discriminant functions for splice sites, 5'-coding, internal exon, and 3'-coding region recognition have been developed. A gene structure prediction system **FGENE** has been developed based on the exon recognition functions. We compute a graph of mutual compatibility of different exons and present a gene structure models as paths of this directed acyclic graph. For an optimal model selection we apply a variant of dynamic programming algorithm to search for the path in the graph with the maximal value of the corresponding discriminant functions. Prediction by **FGENE** for 185 complete human gene sequences has 81% exact exon recognition accuracy and 91% accuracy at the level of individual exon nucleotides with the correlation coefficient (C) equals 0.90. Testing **FGENE** on 35 genes not used in the development of discriminant functions shows 71% accuracy of exact exon prediction and 89% at the nucleotide level (C=0.86). **FGENE** compares very favorably with the other programs currently used to predict protein-coding regions. Analysis of uncharacterized human sequences based on our methods for splice site (**HSPL, RNASPL**), internal exons (**HEXON**), all type of exons (**FEXH**) and human (**FGENEH**) and bacterial (**CDSB**) gene structure prediction and recognition of human and bacterial sequences (**HBR**) (to test a library for E. coli contamination) is available through the University of Houston, Weizmann Institute of Science network server and a **WWW** page of the Human Genome Center at Baylor College of Medicine[1].

1. This work was supported by awards from National Center for Human Genome Research (NIH) and ARCO Foundation to V.V.S.

Introduction

Significant success has been made in coding region identification, however perfect prediction of eukaryotic gene structure continues to be a challenging problem. A test of performance of the latest versions of the most successful gene prediction programs: ***GeneModeler, GeneId, Grail*** and ***GeneParser*** shows that they have an accuracy of exact exon prediction: 2%, 33-42%, 31-52% and 47%, respectively (Snyder & Stormo, 1994), that motivates the development of new approaches to solve this problem.

Most gene prediction systems combine information about functional signals and the regularities of coding and intron regions. On this basis, potential first, internal and terminal exons can be revealed and the top ranking combination of them will present the predicted gene structure. The program ***SORFIND*** (Hutchinson,Hayden,1992) is designed to predict internal exons based on codon usage (around splice sites and in a potential open reading frame) and Berg and 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% (for all 5 confidence levels) or 45% with a specificity of 41% (for the first 3 confidence levels). A dynamic programming approach (alternative to the rule-based approach) was applied by Snyder and Stormo (1993) to internal exon prediction.. It accomplishes an exhaustive and mathematically rigorous search for the globally optimal solution. A sequence is divided into exons and introns by finding the best internally consistent set of high-scoring exon and intron subsequences. Weights for the various classification procedures are determined by training a feed-forward neural network to maximize the number of correct predictions. ***GeneParser*** precisely identifies 76% of internal exons in sequences between the first and last exons, but the structure of only 46% exons

was exactly predicted when tested on entire GenBank entry sequences.

We have developed a program (***HEXON***) for the prediction of internal exons of human genes. The program is based on a splice site prediction algorithm that uses a linear discriminant function to combine information about significant triplet frequencies of various functional parts of splice site regions and preferences of oligonucleotides in protein coding and intron regions (Solovyev, Lawrence, 1993a; Solovyev, Salamov, Lawrence, 1994a). For exon prediction, we combine in a linear discriminant function 5 characteristics describing the 5'-intron region, donor splice site, coding region, acceptor splice site and 3'-intron region for each open reading frame flanked by GT and AG base pairs. The accuracy of precise internal exon recognition on a test set is 77% with a specificity of 79%. The recognition quality computed at the level of individual nucleotides is 89% for exon sequences and 98% for intron sequences. ***HEXON*** has a better exact exon prediction accuracy than other internal exon prediction programs, but it is trained and tested on internal gene regions (including intron and internal exon sequences) and may be useful for analysis of partially sequenced genes.

To predict 5'- and 3'-flanking coding exons, the ***FEXH*** (find human exons) program has been developed (Solovyev, Salamov, Lawrence, 1994b). A discriminant function for 5'-exon prediction consists of hexanucleotide composition of upstream region, triplet composition around the ATG codon, ORF coding potential, donor splice site potential and composition of the downstream intron region. A discriminant function for 3'-exon prediction included octanucleotide composition of upstream intron region, triplet composition around the stop codon, ORF coding potential, acceptor splice site potential and hexanucleotide composition of downstream region. We united 5'-, internal and 3'-discriminant functions in exon predicting program ***FEXH***. ***FEXH*** exactly predicts 70% of 1016 exons from the set of 181 complete genes with specificity 73%, and 89% exons are exactly or partially predicted. On the average, 85% of nucleotides were predicted accurately with specificity 91%. Although ***FEXH*** compares favorable with the other programs currently used to predict coding exons (see review in Snyder & Stormo, 1993,1994), this program shows only the positions of candidate exons and does not attempt to produce assembled genes.

To date, ***GeneModeler*** (Fields and Soderlund,1990), ***GeneID*** (Guigo et al.,1992) and ***XGRAIL*** (Xu et al., 1994) are integrated packages that predict gene structure from genomic DNA. The first two methods rely on revealing the functional motifs such as start and stop codons, splice sites and poly(A) signals; then on sequential filtering evaluation of the assembled combination of a gene component. ***GeneID*** can precisely identify 54% of real exons with correct splice boundaries (Guigo et al.,1992). ***XGRAIL*** is the most widely used of the coding sequence identification program. It based on neural network recognizers of exon candidates, filtering them by a set of heuristic rules and uses dynamic programming approach to build a gene models (Xu et al.,1994). Dynamic programming is used 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. The result of optimization by dynamic programming and gene structure prediction quality are clearly dependent on exon recognition and optimizational functions.

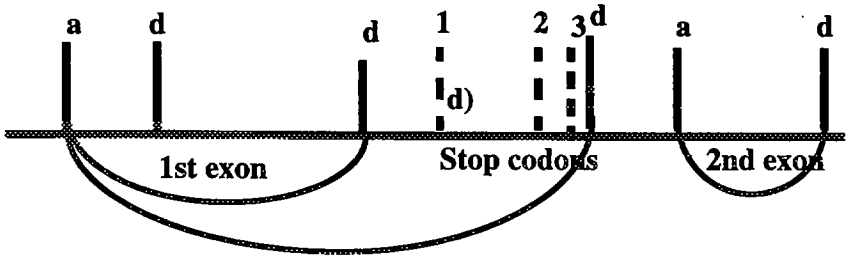
As mentioned above, an accuracy of exact exon identification by gene prediction programs (including ***XGRAIL***) is less than or about 50% when they were tested on the same data set of human genes (Snyder,Stormo,1994). This stimulated us to develop gene structure prediction algorithm using our improved splice site and exon recognition functions (Solovyev, Salamov, Lawrence, 1994a,b). In the present work we combine linear discriminant analysis and dynamic programming approaches in our gene prediction system using the advantages of natural functional description of the former and fast rigorous searching scheme of the later one (Solovyev, Lawrence, 1993b).

Materials and Methods

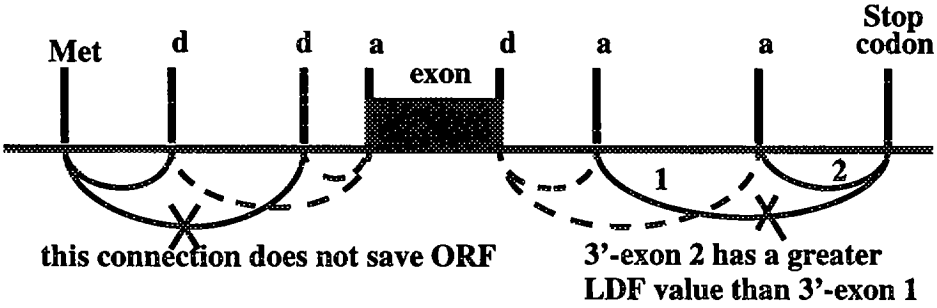
The data sets

We have taken all entries including complete human gene sequences from GenBank (release 81,1994) (Cinkosky et al., 1991). Some entries were divided if they had several genes in them. When a gene had alternative splicing variants, the first of them was selected. Entries with stop codons in-frame of annotated coding regions and nonstandard splice site conservative dinucleotides were discarded, due to possible errors in their description. The size of the resulting set was 185 genes. The set has been used to compute parameters of 5' and 3'-exon discriminant functions and thresholds of the main algorithm. A similar procedure was followed, this time using Gen-

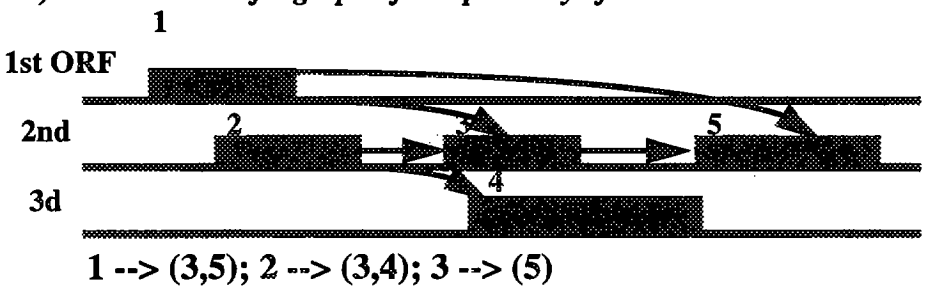
a) Search for a group of potential internal exons



b) Selection of 5'- and 3'-coding exons for each internal exon



c) Construction of a graph of compatibility of internal exons



d) Search for the path with maximal weight in the directed acyclic graph by dynamic programming

The weights are the corresponding LDF values of predicted exons

Fig.1. Construction of gene models by dynamic programming approach.

Bank 84 release. A further 35 entries (having 1994 in their LOCUS description line) with complete genes were thus identified, which we used to evaluate the performance of the ***FGENE*** program.

The algorithm

A general scheme of our gene prediction algorithm is shown in Figure 1 and realized in ***FGENE*** program. We build ***FGENE*** based on linear discriminant functions of internal (LDF_{in}) (Solovyev, Salamov, Lawrence, 1994a), 5'- coding, and 3'-coding ($LDF_{5'}$ and $LDF_{3'}$) (Solovyev, Salamov, Lawrence, 1994b) exon recognition which were described earlier.

We will explain the main steps of the algorithm (Fig.1) and demonstrate them for an example of prediction for GenBank entry **HUMIL8A** (length - 5191bp). The gene from this entry has 4 exons localized in the following positions:(1) 1584 - 1647; (2) 2464 - 2599; (3) 2871 - 2954 and (4) 3370 - 3385.

1. IDENTIFYING POTENTIAL INTERNAL EXONS

First, we create a set of potential internal exons (Figure 1a) using the linear discriminant function (LDF_{in}) for calculating a score of any ORF (open reading frame) between any pair of conservative splice site dinucleotides *AG* and *GT*. We use the threshold score 4.0 for accepting of potential exon, which is less than the optimal for exon prediction without assembling them. Further assembling will remove many of overpredicted exon variants.In the HUMIL8A sequence, 12 potential internal exons were selected. Their positions and LDF values are given in Table 1.

2. RECOGNITION OF POTENTIAL 5'-CODING REGIONS

For each of the predicted potential internal exon, we search for possible 5'-coding region with maximal LDF($Z_{5'}$) value (Figure 1b). We consider all sequence regions having donor site conservative dinucleotide *GT* located more than 60 bp on the left of the acceptor site of the corresponding internal exon and having a compatible ORF starting with Met codon. From these sequences we analyze all

TABLE 1. *Predicted potential exons for GenBank entry HUMIL8A.*

N	Location of	LDF_{in}	ORF	optimal 5'-exon	optimal 3'-exon
	internal exons	LDF	#	location(LDF)	location(LDF)
1	1618 - 1647	6.1	3		3253-3275 (7.9)
2	2464 - 2599	9.5	3	1584-1647 (10.5)	3370-3385 (11.6)
3	2871 - 2954	14.3	1		3370-3385 (11.6)
4	2878 - 2954	13.9	1		3370-3385 (11.6)
5	2892 - 2954	10.2	3		4735-4797 (4.0)
6	2892 - 2954	10.0	1		3370 3385 (11.6)
7	2894 - 2954	9.8	3		4735-4797 (4.0)
8	2894 - 2954	9.7	1	1584-1647 (10.5)	3370-3385 (11.6)
9	2896 - 2954	6.7	1		3370-3385 (11.6)
10	2896 - 2954	7.0	3	1584-1647 (10.5)	4735-4797 (4.0)
11	2914 - 2954	5.6	1		3370-3385 (11.6)
12	2914 - 2954	5.4	3	1584-1647 (10.5)	4735-4797 (4.0)

variants having the discriminant function ($Z_{5'}$) value more than 0.

For HUMIL8A sequence only one potential 5' exon with location 1584-1647 and value of ***LDF*** $Z_{5'}$=10.5 was predicted. In this example only exons 2,8,9 and 12 are compatible with the predicted 5'-exon 1584-1647 (Table 1). It must be noted that not each internal exon will have a 5'-exon after this procedure.

3. *Recognition of potential 3'-coding regions.*

For each of the potential internal exons, we search for possible 3'-coding region with maximal LDF($Z_{3'}$) value (Figure 2b). We consider all sequence regions having acceptor site conservative dinucleotide ***AG*** located more than 60 bp on the right of the donor site of the corresponding internal exon and having a compatible ORF with Stop codon at the end. From these sequences we analyze all variants having the discriminant function ($Z_{3'}$) value more than 0.

For HUMIL8A sequence 47 potential 3'-coding exons were predicted; for example: (3370-3385) with $Z_{3'}$=11.6, (4735-4797) with $Z_{3'}$=4.0, etc. All 12 internal exons have corresponding 3'-exons. For example 1st internal exon (1618-1647) compatible with 3'-exon (3253-3275) with corresponding value $Z_{3'}$=7.9. Selected 3'-coding regions are given in Table 1.

4. *Construction of gene models*

Each predicted exon can be characterized by: (1) the corresponding value of LDF; (2) an assigned reading frame rf {1,2,3}. From the set of predicted internal, 5' and 3'- exons it is possible to construct a model of gene structure for a given sequence. Note that particular potential 5'- and 3'-exons are assigned to each internal exon, therefore we will consider internal exon combinations as potential gene models. The case of gene structure without internal exons will be treated especially.

Assume that we have array of internal exons ordered according with their start positions in a analyzed sequence. Let us define the term of ***compatible exons***. Any two (i-th and j-th) predicted exons (i > j) are considered compatible if: (1) j-th exon localized downstream and the distance between the end of exon i and the beginning of exon j is more than the minimum intron length (60 bp); 2) ORF of these exons are compatible upon their merging, i.e. after removing corresponding intron sequence ORF of the i-th exon proceeds to ORF of the j-th exon and no in-frame stop codons are observed.

We denote the graphical representation of compatibility of exons as an *exon compatibility graph* (***ECG***). The nodes of ***ECG*** are internal exons and edges links compatible ones (Figure 1c). Note, that according to its definition ***ECG*** is a directed and acyclic graph. This graph may be presented by the list form:

1: 3,5
2:4,5
3:5
4:
5:

The first exon is compatible with exons number 3 and 5; the second exon - 4, 5.; the third exon - 5. The exons 4-th and 5-th have no compatible internal exons.

Because of any gene model must consist of a subset of compatible exons, we have to select this subset from all possible groups of such exons. These groups may be presented as paths in compatibility graph.

Each path in this graph going through compatible nodes can be characterized by a weight, which is selected in this algorithm version as the sum of LDF values of its constituent exons. If there is 5'-exon compatible with the first internal exon of the path (first node) and 3'-exon compatible with the last internal exon (last node), they are considered as the first and last nodes of the given path and their LDF values added to the weight of path. For example, for path going through internal exons e_1, e_2, ..., e_n, the total weight of the gene model is equal to:

$$W = \sum_{i=1}^{n} Z_{in}(e_i) + Z_5(e_0) + Z_3(e_{n+1})$$

where Zin, Z_5, Z_3 - LDF values of internal, 5' and 3'-exons selected for e_1 and e_n exons, respectively. $Z_5(e_0) = 0$, if the first internal exon e_1 have no compatible 5'-exon and $Z_3(e_{n+1}) = 0$, if internal exon e_n have no compatible 3'-exon. Our goal to find a path with the maximal W value among all possible paths through compatible exons.

This optimization problem can be solved using a dynamic programming approach to search for an optimal path in acyclic directed graph (Solovyev, Lawrence,1993). We modified one such algorithm, described in (Cormen, Leiserson, Rivest, 1990), and implemented it in ***FGENE*** program.

Let the number of preselected internal exons is N; $MW(i)$ is the maximal weight subpath to node i, consisted of i and nodes predicessing i in ***ECG***; $P(i)$ is the predicessing i node of this subpath; K_i is the number of exons compartible with exon i.

Initially we assign: $P(i)=0$ and $MW(i)=LDF_{in}(e_i)+LDF_5(e_i)$. Then, for each node i, starting from 1, we check all his compatible K_i exons $(j=1,...,K_i)$ and assign $MW(j)=MW(i)+LDF(e_j)$, if $MW(i)+LDF(e_j) > MW(j)$. Also, in this case $P(j)=i$. Finally, the path with the maximal weight will be finishing in node m, which has maximal value of $MW(i)+LDF_3(e_i)$, $(i=1,...,N)$. Using $P(i)$ we restore the set of maximal path exons, by begining from the terminal m. The number of exon (j) predicessing exon (i) is $j=P(i)$.

The running time of this algorithm is O(n+m), where n - number of nodes (potential internal exons) and m - number of edges (total number of all links between compatible exons) of the ***ECG*** graph.

For our illustrative example we obtained a list of compatible downstream internal exons, for each internal exon from Table 1:

1 : (2, 8, 10, 12)
2 : (3, 6, 7).

The first exon is compatible with exon number 2, 8, 10 and 12; the second exon - 2, 6 and 7. Exons 3-12 have no compatible internal exons.

In the HUMIL8A sequence example a path with maximal weight goes through internal exons 2 (2464 - 2599) and 3 (2871-2954). Exon 2 has a corresponding 5'-exon (1584-1647) with $LDF_{5'}$=10.5 and exon 3 has corresponding 3'-exon (3370-3385) with $LDF_{3'}$=11.6. Therefore path through internal exons 2 and 3 has the following weight: W=10.5 + 9.5+14.3+11.6=45.9.

Thus, for HUMIL8A entry all components of annotated gene structure is precisely predicted:

1. 1584 - 1647 (5'-coding region)
2. 2464 - 2599 (the first internal exon)
3. 2871 - 2954 (the second internal exon)
4. 3370 - 3385 (3'-coding region).

For estimation of an algorithm performance we will use the following measures (Fickett and Tung, 1993; Snyder, Stormo, 1993; 1994; Dong, Searls, 1994). Sensitivity (S_n) measures the fraction of the true examples that are correctly predicted and specificity (S_p) measures the fraction of the predicted examples that are correct:

$$Sn = \frac{TP}{TP + FN}; \qquad Sp = \frac{TP}{TP + FP}$$

True positives (TP) is the number of correctly predicted and false positives (FP) is the number of falsely predicted authentic splice site positions (or exons); true negatives (TN) is the number of correctly excluded and false negative (FN) is the number of falsely excluded pseudosite positions (or pseudoexons).

Correlation coefficient (C) is an important accuracy criterion that takes the relation between correctly predictive positives and negatives as well as false positives and negatives into account (Matthews 1975):

$$C = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)\,(FP + TN)\,(TP + FN)\,(TN + FN)}}$$

Results and discussion

Based on the foregoing discriminant functions two coding region prediction algorithm have been developed. The ***FGENEH*** algorithm (described above) builds gene model based on searching an optimal combination of compatible exons. In addition, we have developed a computer program ***FEXH*** (Solovyev, Salamov, Lawrence, 1994b), which predicts a set of coding exons in a given sequence without assembling them. The program initially predicts internal exons based on internal exon discriminant function. Then it searches for 5'-coding region starting from the beginning of the sequence until the end of the first predicted internal exon. In this region the 5'-coding exon with the maximal weight of the first exon discriminant function is selected. After that we search for 3'-coding region starting from the beginning of the last predicted internal exon until the end of the sequence. In this region the 3'-coding exon with the maximal weight of the last exon discriminant function is selected. ***FEXH*** is less sensitive to underprediction of an exon, that can significantly distort the obtained gene model.

We analyze the performance of ***FEXH*** and ***FGENEH*** on training (185 genes) and new test (35 genes) sets of human entry sequences. The results of prediction are presented in Table 2.

At the level of complete exon sequences, ***FEXH*** precisely identifies 611 of 755 (81%) internal exons on the training and 116 of 168 (69%) on the test set. The accuracy of 5' and 3' exons is significantly less, than internal. It exactly predicts 17 of 35 (49%) 5'-exons and 19 of 35 (54%) 3'-exons from the test set sequences. At the nucleotide level S_n=85%, S_p=79% and C=0.8 for the test exon sequences.

Some of predicted pseudoexon ORFs are removed in a gene structure predictive system ***FGENEH*** due to assembling procedure, that improves the accuracy of prediction. It precisely identifies 916 of 1125 (81%) exons from the training set and 168 of 238 (71%) exons from the test set sequences. At the nucleotide level S_n=89%, S_p=88% and C=0.86 for the test entry sequences. We have compared the performance of ***FGENEH*** with the results of ***GRAIL-2*** and ***XGRAIL*** programs. ***GRAIL*** is the most widely used of the coding sequence identification program and shows better or similar quality on entry sequences (compare to ***GeneModeler, GeneId, GeneParser*** or ***GenLang*** algorithms) for different test data (Snyder, Stormo,1994; Dong, Searls,1994)). ***GRAIL-2*** can be used via Email server to test the whole gene set and ***XGRAIL*** can construct a gene model by "optimal" assembling some of predicted exons.

We can see that the current version of ***FGENE*** has accuracy much better than ***GRAIL-2***, especially in exact exon prediction (~ 20% more accurate). The higher accuracy is probably based on quality of our splice sites recognizers, which are more accurate than the others known.

The ***GRAIL X***-Windows (***XGRAIL***) version with model construction has been applied to the prediction of gene structure of the set of new 35 genes. Their sequences were not used in ***FGENE*** as probably well as in ***XGRAIL*** training. We select the best ***XGRAIL*** predicted model for the accuracy estimation. The average prediction results are presented in Table 2. We can see that the current version of ***FGENE*** predicts exactly 71% of 238 real exons, but ***XGRAIL*** the best model predicts only ***48%*** ones. Results of prediction for each test gene are given in table 3. This results show the ***XGRAIL*** accuracy similar with reported by (Snyder and Stormo, 1994).

Thus, our approach has better performance on new test sequences. Moreover, we do not use special heuristic rules for exon selection in this variant of gene prediction, in distinction with the GRAIL algorithm. It must be noted, that higher accuracy of coding region prediction might be reached if ORFs of analyzing sequence have some similarity with known protein sequences. In such case it is very fruitful to use this information during exon selection (Gish,States,1993), but present consideration devotes primarily to prediction of new gene structures, which have no significant similarity on nucleic as well as protein level in the current data bases.

One of the hardest problem of gene prediction is identifying 30-60 bp exons in 100000 bp of bulk DNA. ***FGENE*** performs well on such examples. There are 13 GenBank entry sequences from our test set that have less than 9% of the sequence occupied by protein coding regions. For example, entry ***HSZNGP1*** 9823 bp has 9% coding sequences, and ***FGENE*** predicts all its exons; entry ***HUMFMR1S*** 61613 bp has only 3% coding DNA, and ***FGENE*** predicts exactly 14 of its 17 exons.

Computational aspects

FGENEH is implemented in Fortran 77 and runs on Sun Sparc and DEC Alpha workstations. ***FGENEH*** is quite fast for analysis of usual gene size sequences (5000 -10000 bp). Searching average gene structure in them takes about 1 min on an Sun SparcClassic workstation. However, analysis of sequence of about 200000 bp takes 10 min on

TABLE 2. The performance of FEXH, FGENEH and GRAIL on the training and test sets. Sn^{exon} and Sp^{exon} are the sensitivity and specificity of exact exon prediction; Sn^{nucl} and Sp^{nucl} are the sensitivity and specificity of total exon nucleotide prediction; C-correlation coefficient.

	training entry set					test entry set				
	Sn^{exon}	Sp^{exon}	Sn^{nucl}	Sp^{nucl}	C	Sn^{exon}	Sp^{exon}	Sn^{nucl}	Sp^{nucl}	C
FEXH	72%	62%	88%	80%	0.82	64%	55%	85%	79%	0.80
FGENEH	81%	78%	91%	91%	0.90	71%	67%	89%	88%	0.86
GRAIL2	50%	55%	82%	87%	0.82	45%	68%	71%	88%	0.77
XGRAIL						48%	71%	71%	86%	0.73

DEC alpha computer. The maximal size of an input sequence 200000 is installed in the current version of ***FGENEH***. The size of compatibility graph matrix has to be approximately NxN/3, where N is the number of predicted potential internal exons. Now ***FGENE*** permits N up to 1000 and this is enough for analysis of tested 200000 bp sequence. The search time for predicting potential internal exons grows linearly with sequence length. The memory requirement of the program and run time increase with the square of the number of potential internal exons.

Gene-Finder services

Using the methods presented in our papers we developed a set of sequence analysis programs which are useful for various aspects of gene discovery. The group includes the following programs: ***splice site prediction*** (***HSPL***); recognition of exon-exon junction in cDNA (***RNASPL***), which is useful for ***selecting optimal PCR primers*** in internal exon regions during gene sequencing, when starting with a sequence of cDNA clone; ***internal exons*** (***HEXON***) ***and all type of exons*** (***FEXH***) ***prediction***; ***human gene structure prediction*** (***FGENEH***); ***bacterial gene structure prediction (CDSB)*** and recognition of human and bacterial sequences ***(HBR)*** to ***test a library for E. coli contamination*** by sequencing example clones. Analysis of uncharacterized sequences based on our methods is available through the University of Houston network server by sending the file containing a sequence to ***service@bchs.uh.edu*** with the subject line ***hspl, rnaspl, hexon, fexh or fgeneh or "man fgeneh"*** to receive instructions about the sequence format. During the first month of this server installation (October, 1994) about 600 requests for sequence analysis have been received. These programs are also installed in Weizmann Institute of Science server: ***services@bioinformatics.weizmann.ac.il.***

Gene-Finder programs can be found on ***World Wide Web through the BCM Human Genome Center*** Search launcher Home page (Fig.2) ***URL: http: // kiwi.imgen.bcm.tmc.edu: 8088 /search-launcher /launcher.html*** for access to Gene-Finder and Secondary structure prediction Help files and programs. There were about 1200 times someone ran a gene-finder script through WWW for the first 3 months of 1995.

Bibliography of literature

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

Cinkosky, M.J.; Fickett, J.W.; Gilna, P.; Burks, C. 1991. Electronic Data Publishing and GenBank. *Science* 252: 1273-1277.

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

Dong, S., Searls, D.B. 1994 Gene structure by Linguistic methods. Genomics, 23,540-551.

Gelfand, M.S., Roytberg, M.A. 1993. Prediction of the exon-intron structure by a dynamic programming approach. *BioSystems*, 30, 173-182.

Gish, W., States, D. 1993. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3, 266-272.

Guigo, R.; Knudsen, S.; Drake, N.; Smith, T. 1992. Prediction of gene structure. *J.Mol.Biol.* 226: 141-157.

Table 3.The comparison with the best XGRAIL gene model using 35 test entry sequences presented to GenBank in 1994. E_{ex} is the number of exact predicted exons, C_{nuc} is the correlation coefficient computed based on prediction at the level of individual nucleotides.

GenBank	*Length*	N_{ex}	*XGRAIL*		*FGENEH*	
ENTRY	*nucl.*		E_{ex}	C_{nuc}	E_{ex}	C_{nuc}
HSABLGR1	35962	8	0	0.00	3	0.61
HSCYCLA	8363	8	0	0.44	5	0.92
HSDAO	9903	4	3	0.80	3	0.85
HSHLADMAG	5190	5	2	0.91	1	0.81
HSHLADMBG	6933	6	3	0.86	5	0.97
HSNCAMX1	16288	28	23	0.96	21	0.85
HSU01102	4995	3	1	0.07	2	0.91
HSU04636	9453	10	4	0.69	9	0.85
HSU05259	5670	5	3	0.91	4	0.76
HSUHSU07807	4839	3	0	0.00	2	0.37
HSU08198	2344	7	5	0.98	5	0.94
HSUBR	3321	3	0	0.18	2	0.85
HSZNGP1	9823	4	3	0.88	4	1.00
HUMCSN2A	10608	6	2	0.57	0	0.77
HUMDZA2G	14694	4	2	0.87	3	0.99
HUMFMR1S	61613	17	4	0.56	14	0.86
HUMGAD45A	5378	4	2	0.79	3	0.84
HUMGCK	7807	10	8	0.96	8	0.94
HUMHPARS1	11551	7	5	0.97	5	0.86
HUMIBP3	10884	4	4	1.00	4	1.00
HUMIGERA	7659	5	1	0.71	3	0.71
HUMLHDC	32351	12	5	0.85	8	0.85
HUMMCHEMP	2776	3	3	1.00	3	1.00
HUMMET2	1703	3	1	0.84	3	0.84
HUMMGPA	7734	4	1	0.70	1	0.60
HUMMHCP42	5141	10	9	0.99	10	1.00
HUMMHHLA	3810	7	3	0.84	3	0.85
HUMOSTP	10881	6	3	0.90	5	0.95
HUMPBGDA	10024	14	4	0.67	9	0.82
HUMPLA	2967	5	5	1.00	5	1.00
HUMPYYP1	1935	3	1	0.94	1	0.79
HUMREGHOM	3411	5	1	0.62	4	0.97
HUMRPIB2	4337	5	1	0.63	4	0.97
HUMTBGA	8769	4	1	0.79	3	0.85
HUMTPALBU	6172	6	2	0.70	3	0.68
Average:		*238*	*48%*	*0.67*	*71%*	*0.86*

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

Fields, C.; Soderlund, C.A. 1990. gm:a practical tool for automating DNA sequence analysis. *CABIOS* 6: 263-270.

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

Mathews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem.Biophys.Acta* 405: 442-451.

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.

Snyder, E.E.,Stormo, G.D. 1994, Identifying genes in genomic DNA sequences. To appear in *Nucleic Acid and Protein sequence analysis: A practical Approach*, Second edition (in press).

Solovyev, V.V., Lawrence, C.B. (1993a) Identification of Human gene functional regions based on oligonucleotide composition. In *Proceedings of First International conference on Intelligent System for Molecular Biology* (eds. Hunter L., Searls D.,Shalvic J.), Bethesda, 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. 1994a. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* 22, 24, 5156-5163.

Solovyev, V.V., Salamov, A.A., Lawrence C.B. 1994b. The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. In *Proceedings of the Sec-*

ond International Conference on Intelligent Systems for Molecular Biology (eds. Altman R., Brutlag D., Karp P., Lathrop R., Searls D.), Stanford, CA, 354-362.

Xu, Y., Einstain, R.J., Mural ,R., Shah ,M., Uberbacher, E.C. An improved system for exon recognition and gene modeling in Human DNA sequences. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (eds. Altman R.,Brutlag D.,Karp P.,Lathrop R.,Searls D.), Stanford, CA, 376-383.

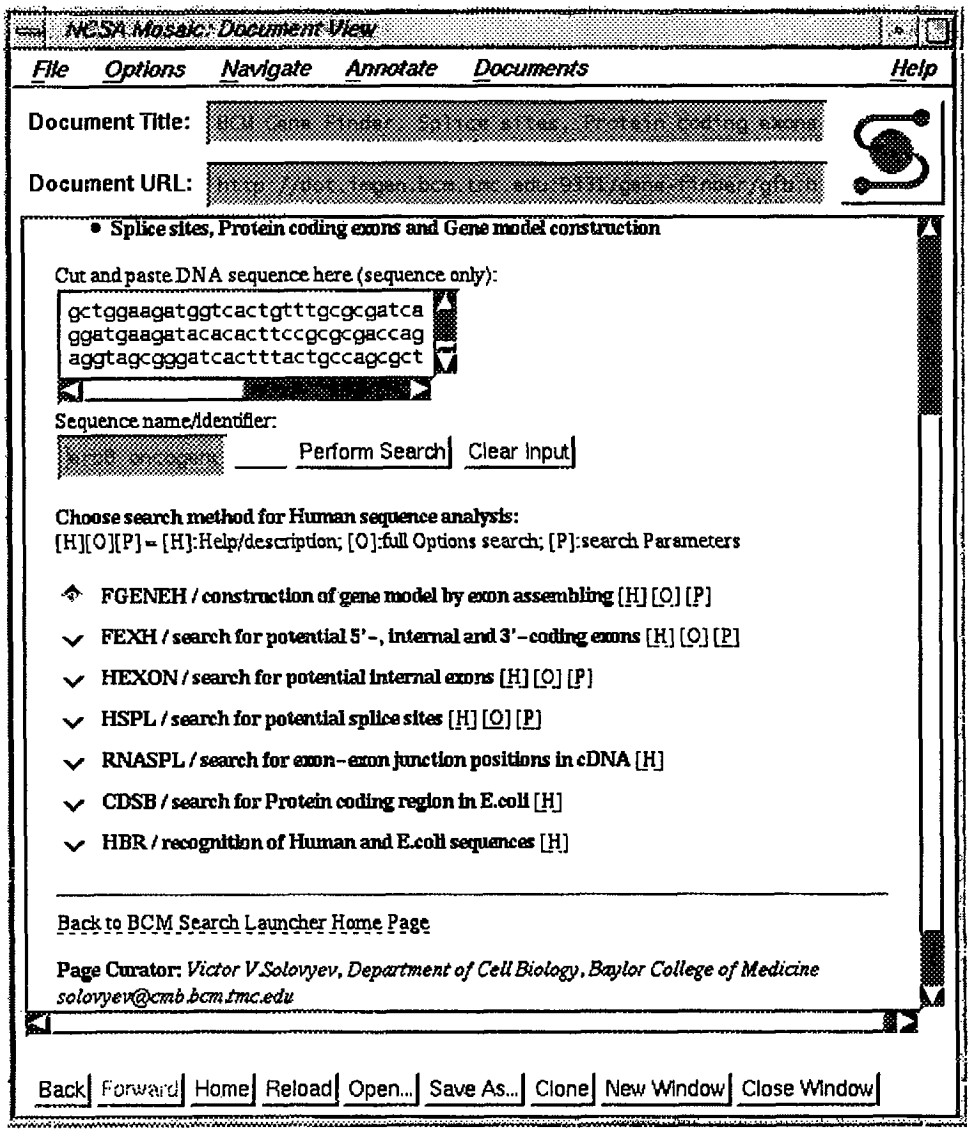


Fig. 2. ***WWW Gene-Finder*** page, where you can past your sequence, run the programs and read the results and help files.