ANALYTICAL BIOCHEMISTRY **265,** 269–274 (1998)
ARTICLE NO. AB982876

Homology-Based Gene Prediction Using Neural Nets

Yudong Cai and Peer Bork
EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany; and Max-Delbruck-Center for Molecular Medicine, Berlin-Buch, Germany

Received June 9, 1998

We have developed and implemented a method for computational gene identification called GIN (gene identification using neural nets and homology information) that has been particularly designed to avoid false positive predictions. It thus predicts 55% of all genes tested correctly, has a specificity of 99%, but also has an overall accuracy of 92% on a benchmark set of 570 vertebrate genes constructed by Burset and Guigo. The method combines homology searches in protein and expressed sequence tag databases with several neural networks designed to recognize start codons, Poly(A) signals, stop codons, and splice sites. Predicted exons are assembled into genes using a homology-based scoring function. GIN is able to recognize multiple genes within genomic DNA as demonstrated by the identification of a globin gene (γ-globin-1(G)) that has not been annotated as a coding region in the widely used the test set of Burset and Guigo. Furthermore, GIN identifies more than 107 other protein hits in noncoding regions and classifies them into possible pseudogenes or splice variants. © 1998 Academic Press

Key Words: **gene prediction; neural nets; homology searches.**

With the scaling up of sequencing in genome projects of higher eukaryotes such as *Caenorhabditis elegans* or human, there is an increasing demand for accurate primary annotation of genomic DNA, including the prediction of the genes and their functions. Although there are complementary data available, such as expressed sequence tags (ESTs)[1] or results of "exon trapping," gene prediction in large amounts of genomic DNA mostly relies on computational methods. Numerous methods have been published for prediction of genes and shown to perform well for known genes (e.g., FGENEH (18), GeneID (11), GeneParser3 (19), GenLang, (7), GRAIL2 (23), and SORFIND (13), but a recent benchmark of existing methods showed an accuracy between 52 and 84% when applying these methods to genomic DNA that had been published after the methods were developped (6). This set of 570 genes has been recently used to benchmark newer methods, but now the set is known and a real performance is difficult to estimate, although methods obviously become more sophisticated and also perform better with more data available. In practice, sequencing groups still often use several complementary methods and assemble the genes manually as there are several parameters such as sensitivity (Sn) and specificity (Sp) (Table 1) that need to be considered and which vary greatly in the numerous programs available (6).

As it is very important to avoid false positive predictions and to "polluting" sequence databases with wrongly assembled genes, we intended to develop a method with high specificity and a low error rate. Errors can propagate in databases and hamper considerably subsequent analysis. Retraction or updating of erronious database entries is the exception rather than the rule (2).

The challenge in computational gene prediction is the efficient combination of various complementory weak signals that leads to the identification and correct assembly of exons. As we consider homology information to be the strongest signal, we developped a method that tries to extract as much information on the coding region as possible out of a detected similarity with database proteins or ESTs. Homology information has been recently incorporated into several gene prediction programs (e.g., GeneID+ (11), GeneParser3 (19), PROCRUSTE (9), and GenView (16), but usually at the end of the prediction procedure. Here we start with exploiting this signal and refine it using several neural nets for start codons, Poly(A) signals, stop codons, and splice sites (Fig. 1). The most similar approach is PROCRUSTE, which is also using BLASTX

[1] Abbreviations used: EST, expressed sequence tag; ORF, open reading frame; GIN, gene identification using neural nets and homology information; HSPs, high-scoring segment pairs.

TABLE 1

Performance of the Programs in the Test Set of 570 Vertebrate Gene Sequences Constructed by Burset and Guigo (6)

	Accuracy per nucleotide				Accuracy per exon				
Method	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
GIN	0.92	0.99	0.92	0.91	0.78	0.80	0.79	0.11	0.04
GENSCAN	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
GeneID+	0.91	0.91	0.88	0.88	0.73	0.70	0.71	0.07	0.13
GeneParser3	0.86	0.91	0.86	0.85	0.56	0.58	0.57	0.14	0.09
GeneParser2	0.66	0.79	0.67	0.65	0.35	0.39	0.37	0.29	0.17
Genie	0.76	0.77	0.72		0.55	0.48	0.51	0.17	0.33
FGENEH	0.77	0.88	0.78	0.80	0.61	0.64	0.64	0.15	0.12
GRAIL2	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11
GeneID	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
GenLang	0.72	0.79	0.69	0.71	0.51	0.52	0.52	0.21	0.22
SORFIND	0.71	0.85	0.73	0.72	0.42	0.47	0.45	0.24	0.14
Xpound	0.61	0.87	0.68	0.69	0.15	0.18	0.17	0.33	0.13

Note. Sn/nucleotide, TP/(TP + FN); Sp/nucleotide, TP/(TP + FP); CC, (TP*TN − FP*FN)/sqrt(PP*PN*AP*AN); AC, 0.5 * (TP/AP + TP/PP + TN/AN + TN/PN) − 1. AP, actual positives; AN, actual negatives; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives. Sn/exon, (number of correct exons)/(number of actual exons); Sp/exon, (number of correct exons)/(number of predicted exons); Avg., (Sn/exon + Sp/exon)/2; ME, (number of missing exons)/(number of actual exons); WE, (number of wrong exons)/(number of predicted exons).

(10) for fast identification of exons, but the details such as recognition of DNA signals, exon assembly, etc. vary greatly.

When analyzing the matches of putative exons to protein databases, we were surprised to find numerous similarities to proteins in the annotated introns or intergenic region of a test set collected by Burset and Guigo (6). Although most of these hits comprise pseudogenes or even artificial open reading frames (ORFs) stored in public protein databases, the database searches are also indicating distinct splice variants and a real gene that has not been annotated and does not appear in any report on novel gene prediction methods that have used benchmark of Burset and Guigo (6) (e.g., Refs. 5, 12, 15). This raises questions as to the validation of prediction methods as the annotation cannot be taken as the absolute truth.

In the following, we briefly outline our strategy for gene prediction and then discuss the results.

As the expectation of finding for a given query homologous sequences in public databases is increasing (3) and a significant similarity to a protein is the strongest signal for a coding potential, we first scan (after filtering out ALU and B2 DNA repeats using BLASTN program for the public repetitive sequences database (Ref. 14; see Fig. 1)) public protein databases using the BLASTX program (10). As the memory requirements increase with the length of the genomic query sequence, we chop long queries into shorter segments and reassemble them. If no putative exon is indicated, we use the TBLASTX program to compare the genomic query sequence with public available ESTs.

We then combine this information with searches for various DNA signals such as start codon, Poly(A) signal, stop codon, splice site, and branch point signal by using the back-propagation neural network model (21) to determine all the possible exons. The back-propagation model has a multilayered sensory structure and a strong ability for self-organizing and self-adaptability, by learning some representative examples, to know fundamental characteristics of the objects. The self-learning algorithm of the back-propagation model is an iterative procedure. At first, a set of initial weights of the network is given, and then one by one samples are input to the network and the output is calculated. The difference between the calculated value and the expected value is used to update its weights so that the difference can be reduced. This updating process will be repeated until the difference is smaller than a specified error value. After the neural network is trained in the self-learning way by sufficient samples, the final weights are its correct interior representation.

In this research, the four bases of DNA sequences are coded as 4D vectors composed of only 0 and 1 (A, 1000; T, 0100; G, 0010; C, 0001), which are taken as the input of the neural network. Therefore, the number of units in the input layer is equal to the number of nucleotides contained in training samples × 4. The output layer contains one unit. The neural network model just has one hidden layer which contains eight units.

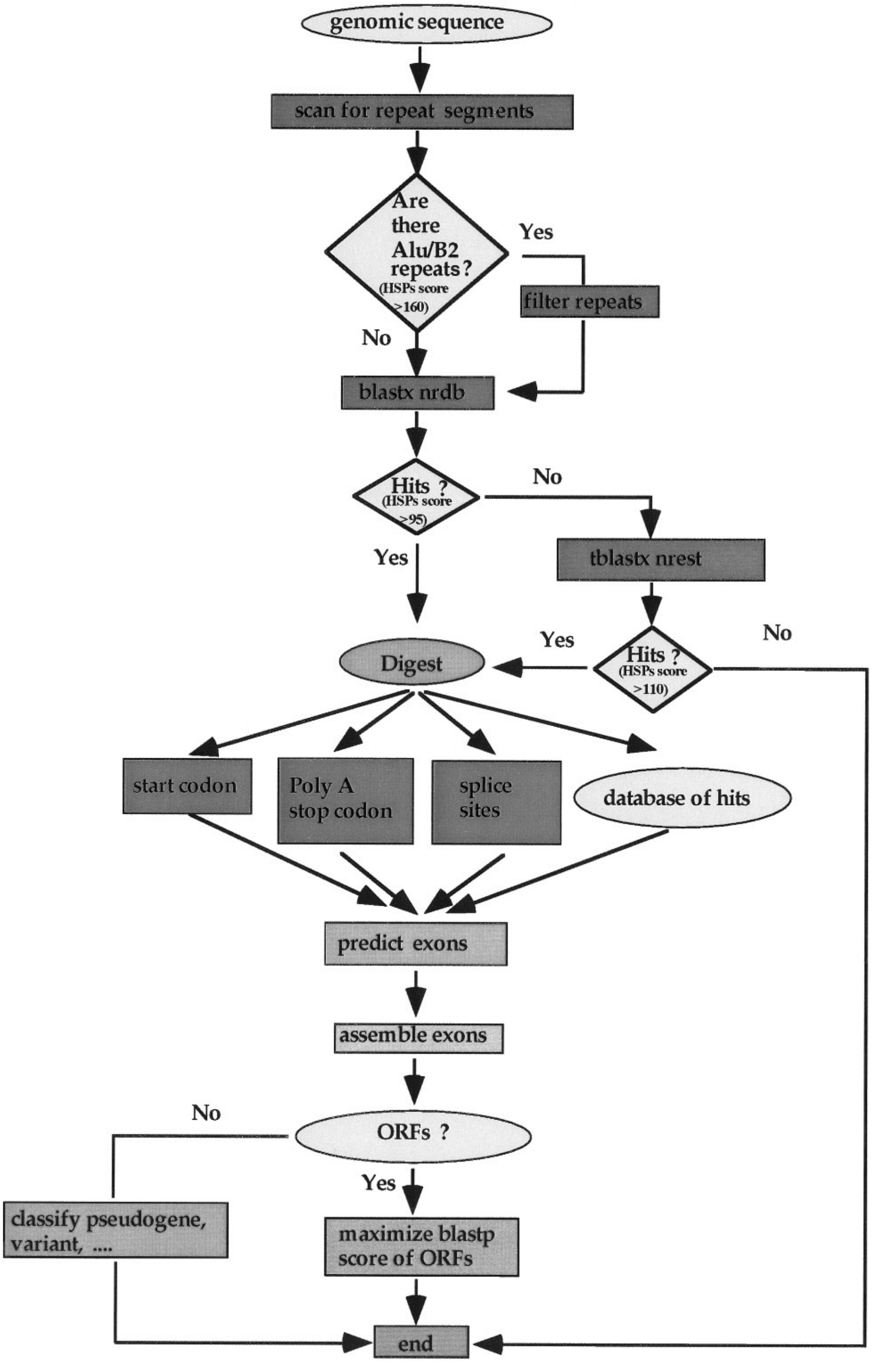

FIG. 1. Flow chart of the GIN system.

The training database includes 420 sequences taken from five eukaryotic organisms: human, rat, cattle, sheep, and monkey. They are all different from the 570 sequences of the ALLSEQ database (6) which was taken to be our test set. Because of the size of the training database, it is not given in this paper, but it is available from the authors upon request.

Let (*i*, *j*) be the exon interval found by sequence signal searches and (*x*, *y*) be the homology hits.

(i, j) is a putitive exon determined by the following rule:

$$|x - i| < 80 \quad \text{and} \quad |y - j| < 80.$$

The remaining part is to find an optimal path through all the possible exon combinations from likely start to stop codons. Our scoring function for this computer-intensive task is again provided by homology information as the highest scoring hit to a matching protein is considered as the most likely gene.

The detailed procedure is the following (Fig. 1):

Step 1. Input a genomic sequence.

Step 2. BLASTN search: sequence → repeat database. Get hits to the best homology repetitive sequence: high-scoring segment pairs (HSPs) score >160.

Step 3. BLASTX search: sequence → NRDB. Get hits to the best homology protein: HSPs score >95.

Step 4. If there are not any hits (HSPs >95) found, go to Step 5; otherwise, go to Step 6.

Step 5. TBLASTX search: sequence → NREST. Get hits to the best homology cDNA sequence: HSPs score >110.

Step 6. Start codon (CCA/GCC|ATGG), Poly(A) signal (AATAAA), stop codon (TAA, TGA, TGA), and splice site (5′ ss, A/CAG|GTRAGT; 3′ ss, YYYYYYYYYYYAG; branch point signal, CTRAY; R, purine; Y, pyrimidine) searches by neural network method (back-propagation model).

Step 7. Determine all the possible exons: Let (i, j) be the exon interval found by sequence signal searches and (x, y) be the homology hits. (i, j) is a putitive exon determined by the following rule:

$$|x - i| < 80 \quad \text{and} \quad |y - j| < 80.$$

Step 8. Get all the possible exon combinations from start codon to stop codon. Find all the ORFs by reading the combinations codon by codon until the stop codon (TAG, TAA, TGA) occurs.

Step 9. Predicted gene is the ORF which has the highest similarity score to the best homology protein or cDNA of the genomic sequence.

Significant similarities to proteins that cannot be integrated into gene models are considered as a core unit for a separate gene in order to be able to handle nested genes or simply neighboring genes within a genomic contig. If such core units cannot be assembled into complete gene models, there might be several reasons for it: (i) Gene identification using neural nets and homology information (GIN) is unable to locate DNA signals or misses exons, leading to the conflict of having a likely coding region (determined via a significant homology to a protein); (ii) the protein match might be spurious, for example, due to compositional bias (10); (iii) the protein stored in a public database is an artifact and in fact not coding; (iv) the protein match is caused by a pseudogene; or (v) the match is due to an alternative splice variant.

GIN does not consider those matches as genes (i.e., they are not included in the statistics), but several rules are applied to classify those regions by parsing the outputs of the database search programs. For example, if a "splice variant" is indicated in the header of a matching protein, it is likely that an alternative splicing variant has been found, unless the respective gene is modified in a disease only and thus should not be considered as a good alternative. High-scoring matches that cover only a short part of the matched database protein are considered to be pseudogenes if they cannot be fit into a gene model. Matches to a single protein in the TREMBL database (EMBL entries translated by a script from and provided by Thure Etzold, EBI) or in the PIR fraction that does not have a counterpart in other databases are also treated with caution as we repeatedly identified artificial proteins in this set. All this additional information is provided by GIN.

The performance of GIN was tested on the set of 570 vertebrate multiexon gene sequence (6). The standard measures of predictive accuracy per nucleotide and per exon are shown in Table 1. They indicate that GIN performs considerably better than other methods that utilize homology information (e.g., GeneID+ or GeneParser3), although a direct comparison with PROCRUSTE and GenView is not possible as they do not use the benchmark introduced by Burset and Guigo (6). One difference to PROCRUSTE and GenView that should contribute to an increase in accuracy is the choise of a more exhaustive database. We use an assembly of different databases (SWISSPROT, GENPEPT, TREMBL, and PIR) cleaned for identical sequences using the NRDB script from the NCBI. This daily updated NRDB database probably contains fourfold as many sequences as the annotated SWISSPROT part, although it is still redundant.

Comparison of the accuracy data (Table 1) shows that GIN performs similar to GENSCAN (5), a program that is based on DNA signals, compositional features of exons, introns, and intergenic regions, and that does not use homology information. Only the specificity/nucleotide is currently better than other published methods (for the test set). Another important parameter is how many of the genes are entirely predicted correctly ("gene-level accuracy" as introduced by Burge and Karlin (5)). GIN performs here considerably better than GENSCAN with 0.55 (315 out of 570 genes) for GIN and 0.43 (243 out of 570) for GENSCAN, respectively. The greater accuracy in specificity/nucleotide means that GIN has less false positives in predicting nucleotides; therefore, there is less chance to

TABLE 2

Classification of the 108 Hits in Noncoding Regions of the Test Set

Classes	Number	Intergenic	Intron
Genes	1	1	0
Pseudogenes[a]	71	46	25
Variant	9	2	7
Single high-scoring hit in TREMBL or PIR[b]	27	17	10
Total	108	66	42

[a] Criteria for classifying: High-scoring hits cover only a short part of the matched database protein and they cannot be fit into a gene.

[b] We do not exclude such hits, but caution and manual analysis are recommended.

provide false positives in amino acids in the final predicted protein product. Thus GIN performs better than GENSCAN in gene-level accuracy as mentioned above, although the specificity/exon (specificity in exon level is the proportion of predicted exons that is correctly predicted) is similar to GENSCAN's. However, one has to consider that (i) all of the 570 genes have some homologs in current databases, (ii) the test set is known and thus some of its information is indirectly used for training, and that (iii) the intron/exon structure in the test genes is often rather simple (5).

Furthermore, when analyzing the false positives, we found that a γ-globin-1(G) from monkey (Accession No. X53419) is in fact a correct prediction and a known gene (8) but was not annotated in the benchmark set as coding due to a missing database comment. Thus, we cannot exclude that more genes might be hidden in the test set.

As GIN records all significant similarities of the queries to putative proteins that are likely cores of exons but that cannot be assembled into complete genes, we were able to map and to classify as many as 107 of those hits into the noncoding regions of the 570 query contigs of the test set (Table 2). Some of them are predicted as "alternative splice variants."

These matches provide an additional annotation that is hard to decode into an "accuracy" value. For example, fragments that GIN classifies as "pseudogenes" might in fact be parts of overlooked genes that GIN fails to assemble properly. Also, there were 14 partial hits to reverse transcriptases of transposons which GIN annotates separately. Some of them might in fact be functional.

Taking together (i) the prediction of a complete gene in a region annotated as intergenic, (ii) several indications of alternative splicing, and (iii) several other regions that harbor pseudogenes or genes calls for caution when interpreting percentages of accuracy (Table 1). Thus, the numbers measured on the test set of Burset and Guigo (6) (Table 1) can only be seen as rough guides, and only the praxis will show the real performance.

In addition, the weak point of GIN is rather low prediction power, if no database hits are found. For new genes where we expect approximately 30% without a database hit, the accuracy will go down considerably. And there is also much room for improvement of the GIN system as we currently have no good handle for the prediction of very short exons (<40 nucleotides) that are often overlooked by GIN. Even if we consider all possible exons predicted by DNA signals (a time-consuming step), their occurrence as terminal exons still remains a problem. Another problem remains the extreme compositional bias of some genes. Usually, one can "filter" those regions in the BLAST program suite using, for example, the SEG program (22) in order to avoid spurious hits. Unfortunately, this is at the cost of masking exons, so that a complicated and time-consuming iterative search system has to be applied.

In summary, the GIN system shows that when utilizing homology information in a first step (if applicable) and in combination with signal searches by neural nets, a reasonable accuracy in gene prediction can be achieved. GIN should be complementary to methods such as GENSCAN as its principle is very different. A server for gene prediction using GIN with all the features discussed has been recently provided: http://www.bork.embl-heidelberg.de/fmilpetz/GIN/.

ACKNOWLEDGMENTS

We are grateful to Frank Milpetz for the programming of the WWW server and Yuan Yan-Ping for discussions related to the databases used.

REFERENCES

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
2. Bork, P., and Bairoch, A. (1996) *Trends Genet.* **12,** 425–427.
3. Bork, P., Ouzounis, C., and Sander, C. (1994) From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4,** 393–403.
4. Borodovsky, M., and McIninch, J. (1993) *Comp. Chem.* **17,** 123–134.
5. Burge, C., and Karlin, S. (1997) *J. Mol. Biol.* **268,** 78–94.
6. Burset, M., and Guigo, R. (1996) *Genomics* **34,** 353–367.
7. Dong, S., and Searls, D. B. (1994) *Genomics* **23,** 540–551.
8. Fitch, D., Bailey, W., Tagle, D., Goodman, M., Sieu, L., and Slightom, J. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 7396–7400.
9. Gelfand, M. S., Mironov, A. A., and Pevzner, P. A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9061–9066.
10. Gish, W., and States, D. J. (1993) *Nature Genet.* **3,** 266–272.
11. Guigo, R., Knudsen, S., Drake, N., and Smith, T. (1992) *J. Mol. Biol.* **226,** 141–157.
12. Henderson, J., Salzberg, S., and Fasman, K. (1997) *J. Comp. Biol.* **4**(2), 119–126.

13. Hutchinson, G. B., and Hayden, M. R. (1992) *Nucleic Acids Res.* **20,** 3453–3462.

14. Jurka, J. (1995) Database of Repetitive Elements (Repbase). NCBI Database Repository (ftp://ncbi.nlm.nih.gov/repository/repbase/).

15. Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H. (1996) A Generalized Hidden Markov Model for the Recognition of Human genes in DNA, ISMB-96, AAAI Press, Menlo Park, CA.

16. Rogozin, I., Milanesi, L., and Kolchanov, A. (1996) *Comp. Appl. Biol. Sci.* **12**(3), 161–170.

17. Salzberg, S., Delcher, A., Fasman, K., and Henderson, J. (1997) A Decision Tree System for Finding Genes in DNA. Technical Report 1997-03, Department of Computer Science, Johns Hopkins University.

18. Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1994) *Nucleic Acids Res.* **22,** 5156–5163.

19. Snyder, E. E., and Stormo, G. D. (1995) *J. Mol. Biol.* **248,** 1–18.

20. Warren, P., and Sally, A. (1996) *Mol. Microbiol.* **21**(2), 213–219.

21. Werbos, P. (1974) Beyond Regression: New Tools for Prediction and Analysis in Behavioral Sciences, Ph.D. dissertation, Appl. Math., Harvard University, Cambridge, MA.

22. Wootton, J., and Federhen, S. (1993) *Comp. Chem.* **17,** 149–163.

23. Xu, Y., Einstein, J. R., Mural, R. J., Shah, M., and Uberbacher, E. C. (1994) An Improved System for Exon Recognition and Gene Modeling in Human DNA Sequences, ISMB-94, pp. 376–383, AAAI Press, Menlo Park, CA.

24. Zhang, M. Q. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 559–564.