GENOMICS **51,** 332–339 (1998)
ARTICLE NO. GE985251

Performance-Guarantee Gene Predictions via Spliced Alignment

Andrey A. Mironov,*,[1] Michael A. Roytberg,† Pavel A. Pevzner,‡,[2] and Mikhail S. Gelfand§

*Laboratory of Mathematical Methods, National Center for Biotechnology NIIGENETIKA, Moscow 113545, Russia; †Institute of Mathematical Problems of Biology and §Institute of Protein Research, Russian Academy of Sciences, Puschino, Moscow region 142292, Russia; and ‡Departments of Mathematics and Computer Science, University of Southern California, Los Angeles, California 90089-1113

Received March 5, 1997; accepted January 29, 1998

An important and still unsolved problem in gene prediction is designing an algorithm that not only predicts genes but estimates the quality of *individual* predictions as well. Since experimental biologists are interested mainly in the reliability of individual predictions (rather than in the average reliability of an algorithm) we attempted to develop a gene recognition algorithm that guarantees a certain quality of predictions. We demonstrate here that the similarity level with a related protein is a reliable quality estimator for the *spliced alignment* approach to gene recognition. We also study the average performance of the spliced alignment algorithm for different targets on a complete set of human genomic sequences with known relatives and demonstrate that the average performance of the method remains high even for very distant targets. Using plant, fungal, and prokaryotic target proteins for recognition of human genes leads to accurate predictions with 95, 93, and 91% correlation coefficient, respectively. For target proteins with similarity score above 60%, not only the average correlation coefficient is very high (97% and up) but also the quality of individual predictions is *guaranteed* to be at least 82%. It indicates that for this level of similarity the worst case performance of the spliced alignment algorithm is better than the average case performance of many statistical gene recognition methods.

INTRODUCTION

The large-scale sequencing projects have motivated the need in a new generation of algorithms for computational gene recognition in long uncharacterized DNA sequences. Recently the traditional statistical approach to recognition of protein-coding genes was supplemented by similarity-based approaches (for technical reviews on computer-assisted functional mapping of DNA sequences see Gelfand, 1995; Fickett, 1996a; introduction for users is in Fickett, 1996b). Similarity search can be used to detect genes (Gish and States, 1993) and, in conjunction with the statistical analysis, to predict exon–intron structure of eukaryote genes. Indeed, similarity to an already known gene can provide additional statistical parameters (Snyder and Stormo, 1995), allow the program to choose between several suboptimal genes (Rogozin *et al.*, 1996), and serve as the main scoring function for candidate protein-coding segments (Hultner *et al.*, 1994). Some existing servers perform database simiarity search for a predicted gene as a standard postprocessing procedure (Uberbacher *et al.*, 1996). These approaches utilize the large amount of previously sequenced DNA and are likely to become the method of choice in the future.

Consistent realization of the similarity-based gene recognition is provided by the *spliced alignment* algorithm implemented in Procrustes software (Gelfand *et al.*, 1996a). The algorithm explores all possible exon assemblies in polynomial time and finds the multiexon structure with the best fit to a related protein. This is the main feature of the algorithm distinguishing it from other programs.

Given a *genomic sequence,* the spliced alignment algorithm first finds *candidate exons.* This can be done by selecting all sequence fragments between potential acceptor and donor sites (i.e., between AG and GU dinucleotides) with further *filtration* of this set (in a way that does not lose the actual exons). The resulting set, of course, can contain many false exons, and currently it is impossible to distinguish all actual exons from this set by statistical methods. Instead, the spliced alignment algorithm explores *all* possible assemblies of potential exons and finds an assembly with the highest similarity to the related *target protein.*

One of the main problems in gene recognition is designing an algorithm that would not only predict genes, but estimate the quality of *individual* predictions as well. The *average* performance of an algorithm can be estimated using correlation between the predicted and the correct gene structure (Burset and Guigo, 1996). However, in a real situation the correct

[1] Current address: AnchorGen, Santa Monica, CA 90403.
[2] To whom correspondence should be addressed. Telephone: (213) 740-2407. E-mail: ppevzner@hto.usc.edu.

0888-7543/98 $25.00

TABLE 1

Distribution of the Local Similarities (Negative Logarithms of BLAST Scores)

Entrez score	<3	3	4	5	6	7	8	9	10–11	12–13	14–15
No. of seq.	0	5	30	22	23	18	14	11	29	26	27

Entrez score	16–17	18–19	20–29	30–39	40–49	50–59	60–99	100+
No. of seq.	17	19	91	74	54	37	95	85

gene structure is unknown, and no algorithm provides an estimate for the quality of obtained gene predictions. At best, some algorithms assign quality indicators to particular exons (Uberbacher and Mural, 1991).

Since experimental biologists are more interested in reliability of individual predictions than in the average performance, it is important to develop gene recognition algorithms with guaranteed prediction quality. Recently Sze and Pevzner (1997) used the quality and certainty of fit of a candidate exon to the respective region in the target protein as an indicator of the prediction quality of individual exons. For human genes predicted by spliced alignment with mammalian targets all exons are guaranteed to be correct in one-third of cases, and at least one exon is guaranteed in half of cases.

The present study analyzes the overall certainty of spliced alignment predictions with both mammalian and more distant targets and provides the estimate and bounds for the correlation coefficient between the predicted and the actual gene given the obtained similarity score. The spliced alignment algorithm is tested on the set of all completely sequenced human genes with a known related protein from another species. We study the dependence between the level of protein similarity and the accuracy of predictions and determine the ranges of guaranteed performance. In particular, we demonstrate that if the similarity score is $60 \pm 5\%$, the average correlation is 95%, whereas the correlation for individual predictions is always higher than 75%; it is higher than 80% in more than 97% of cases and higher than 95% in more than 70% of cases, and the prediction is exact in approximately half of all cases. We assume that predictions with a similarity score exceeding 30% are reliable and consider the remaining predictions tentative. At that, the average correlation of reliable predictions for all groups of eukaryotic targets, including plants and fungi, exceeds 93%, and it is 91% for prokaryotic targets. The accuracy of tentative predictions is lower but still higher than the accuracy of statistical gene recognition algorithms as long as the similarity score is above 20%. We also analyze various filtration procedures and demonstrate that stronger filtration provides better results for distant targets despite the danger of overfiltering true exons.

DATA

All human DNA sequences from GenBank (Benson *et al.*, 1997) and EMBL (Stoesser *et al.*, 1997) containing completely sequenced genes were considered. This was done by automated search for text terms *complete gene* or *complete CDS* in the human divisions of GenBank and EMBL (Spring 1996 releases). This preliminary list was supplemented by sequences from Snyder and Stormo (1995), Burset and Guigo (1996), and Gelfand *et al.* (1996b).

A syntactic check was performed first on feature tables to exclude mRNA genes, incomplete and alternatively spliced genes, sequence fragments containing multiple genes, and entries with errors in feature tables (in-frame stop codons, missing start or stop codons). We also removed genes with introns shorter than 70 nucleotides, which do not occur in human genes (Sharp, 1994) and are an indication of an error in the feature table. Genes having unconventional splicing sites breaking the GU–AG rule also were removed. Such sites occur in less than 1% of human genes (Jackson, 1991).

Target sequences were selected using the *Entrez* browser (Schuler *et al.*, 1996). Fifteen genes having no nonprimate relatives and all histone genes were excluded from the sample. For each gene one highest scoring target protein in each of the following categories was considered: mammals (not primates), birds, cold-blooded vertebrates, insects, other animals, plants, fungi, other eukaryotes, and prokaryotes. Genes having the same highest scoring mammalian relatives were considered homologous. Only the longest genomic sequence fragment from each group of homologues was retained. Distribution of local similarities (negative logarithms of BLAST probabilities as given in *Entrez*) is shown in Table 1.

The resulting sample consists of 256 sequences and is available from the Procrustes WWW site. The average sequence length is approximately 8100 nucleo-

TABLE 2

Distribution of the Length of Genomic DNA

Length (kb)	<5	5–10	10–15	15–20	20–30	30–40	55	180
No. of seq.	126	81	25	8	5	9	1	1

TABLE 3

Distribution of the Number of Exons in Human Genes

No. of exons	1	2	3–5	6–10	11–20	21–30	38	54
No. of seq.	28	26	111	70	16	3	1	1

tides; the longest sequence in the sample exceeds 180,000 nucleotides. The distribution of sequence lengths is shown in Table 2. The number of exons in genes ranges from 1 to 54 (Table 3), and their minimum length is 3 bp for initial exons, 17 bp for internal exons, and 5 bp for terminal exons (Table 4).

METHODS

Filtration. Initially all *initial exons* bounded by a start codon <ATG and a candidate donor site >GT, *internal exons* bounded by an acceptor site AG< and a donor site >GT, *terminal exons* bounded by an acceptor site AG< and a stop codon >TGA, >TAA, or >TAG are considered (< and > denote the left and right boundaries of a coding region, respectively). Note that we use the term *exon* as a synonym for *translated part of an exon*, which is the traditional although biologically incorrect use of this term in computational molecular biology. Internal exons should be longer than 16 nucleotides.

Filtration consists of two weak filters removing clearly abnormal exons and a final filter of adjustable strictness.

The first filter removes exons with weak splicing sites as estimated by positional nucleotide weight matrices (Gelfand *et al.*, 1996b). The threshold is set very low and only two actual acceptor sites are filtered out at this step.

At the second step the genomic sequence is divided into subfragments of length 10 kb with 2.5 kb of overlap. Further filtration is performed independently in each subfragment and the candidate exons are evaluated by a scoring function taking into account strength of the splicing sites and the coding potential (Gelfand *et al.*, 1996b). One thousand highest scoring exons are retained in each subfragment. This filter loses 10 actual exons: 1 initial, 6 internal, and 3 terminal.

Overall, the two preliminary filters decrease the number of candidate exons approximately 15-fold, while losing 12 actual exons in the entire sample. It should be noted that the statistical properties of these exons are so unusual that they will likely be lost by any conventional gene recognition algorithm. At the same time, preliminary filtering sharply decreases the number of candidate exons, making the final filter more robust.

At the main filtration step chains of exons of length 1 through 3 with consistent reading frame (no in-frame stop codons) and introns longer than 70 nucleotides are considered. Each chain is scored by the statistics-based function. Denote the score of a chain Γ by $|\Gamma|$. An exon score is now defined as either

$$P(e) = \sum_{\Gamma \ni e} e^{c|\Gamma|},$$

where c is some fixed constant (*partition function rescoring*), or

$$B(e) = \max_{\Gamma \ni e} |\Gamma|$$

(*best chain rescoring*).

The candidate exons are then ranked in the decreasing order of their scores and the given proportion of exons is retained for the spliced alignment procedure. The ranking is performed independently for initial, internal, and terminal exons. The proportion of these three classes of exons in the filter output is 1:3:1, respectively.

Thus the filtering is controlled by two switches (the maximal number of exons in chains $E = 1$, 2, or 3 and the use of P or B scores) and the *filtration stringency* parameter F. This parameter determines the number of exons dependent on the genomic sequence length. Following preliminary analysis, three values of this parameter have been considered: 1 exon per 14 nucleotides ($F = 14$, weak filtration), 1 exon per 33 nucleotides ($F = 33$, moderate filtration), and 1 exon per 100 nucleotides ($F = 100$ strong filtration). Note that if $E = 1$ all filters coincide.

Table 5 presents the results of comparison of different filters. Filters based on chains of two and three exons ($E = 2$ or $E = 3$) outperform single-exon filters ($E = 1$) in the entire range of filtration stringency.

The best weak filtration mode ($F = 14$) seems to be partition function rescoring (P) with two-exon chains ($E = 2$). Relaxing the filtration parameters further does not recover more than one lost exon, while adding many more false exons.

The optimal mode for moderate and strong filtration ($F = 33$ and $F = 100$) is best structure rescoring (B) with three-exon chains ($E = 3$). These options were fixed for further analysis.

Single-exon genes were considered separately in an analogous manner. The minimum length of such genes was set to 180 nucleotides; one candidate exon was retained per 200 bp of the genomic sequence.

Spliced alignment. UNIX version of the Procrustes was used for the sample processing. The WWW version is available at http://www-hto.usc.edu/software/procrustes.

The spliced alignment score was computed using PAM120 amino acid substitution matrix (Altschul, 1991) with linear gap

TABLE 4

Distribution of the Exon Lengths in 202 Human Genes with Three or More Exons

202 initial exons: average length 155 bp, min. length 3 bp, max. length 3051 bp

Len.	1–5	6–10	11–20	21–30	31–50	51–75	76–100	101–150	151–200	201–300	301–1000	>1000
No.	3	3	11	6	20	47	25	36	15	16	17	3

907 internal exons: average length 139 bp, min. length 17 bp, max. length 885 bp

Len.	1–20	21–25	26–30	31–40	41–50	51–75	76–100	100–125	126–150	151–200	200–300	>300
No.	1	6	4	21	24	93	111	164	159	195	103	26

202 terminal exons: average length 191 bp, min. length 5 bp, max. length 1546 bp

Len.	1–5	6–10	11–20	21–30	31–50	51–75	76–100	101–150	151–200	201–300	301–1000	>1000
No.	1	2	6	7	11	21	19	54	35	16	27	3

TABLE 5

Comparison of Different Filtrations

		F						
E		100	50	33	25	20	14	10	No filter
1	—	117	160	186	202	214	224	239	256
2	P	148	192	210	219	228	239	240	256
2	B	144	173	198	214	229	234	240	256
3	P	138	180	206	221	230	236	241	256
3	B	162	198	218	227	231	234	240	256

Note. In each cell the number of sequences in which no overfiltration occurs is shown. Lanes: type of filtration described by *E*—number of exons in chains; *P* or *B*—resp. partition function rescoring or best structure rescoring. Columns: *F*—stringency of filtration (sequence length divided by the number of candidate exons).

penalties (the preliminary analysis demonstrated that the influence of the matrix on the algorithm performance is minor; other gap scoring schemes were implemented). This score was normalized by division of the score of the (trivial) alignment of the target protein with itself.

The quality of prediction was assessed using the correlation coefficient between the predicted and the actual genes,

$$C = \frac{T_P \cdot T_N - F_P \cdot F_N}{\sqrt{(T_P + F_P) \cdot (T_N + F_N) \cdot (T_P + F_N) \cdot (T_N + F_P)}},$$

where T_P and T_N are the numbers of correctly predicted coding (true positive) and noncoding (true negative) nucleotides, respectively, F_N is the number of missed coding (false negative) nucleotides, and F_P is the number of noncoding nucleotides predicted to be coding (false positive).

RESULTS

The average correlation coefficients for different groups of targets are presented at Table 6. It should be noted, however, that the target group is a very rough indicator of the expected prediction quality, since the mutation rates differ significantly between protein families within the same species. Further, since the targets have been chosen by the BLAST database search (via *Entrez*), many targets have only local similarities with the analyzed genes and thus produce artifacts when the (global) spliced alignment is performed.

A better indication of the expected recognition quality is provided by the normalized spliced alignment score. The scatter plots of the correlation coefficient versus the alignment score (Fig. 1) demonstrate that high prediction quality is guaranteed if the alignment score is high. The same figures feature plots of similarity levels providing 100, 95, 90, and 80% guarantee of obtaining the desired correlation coefficient given the observed alignment score and the plot of the average correlation coefficient for the given alignment score. The plots in Fig. 2 provide an estimate for the proportion of predictions having the correlation exceeding 80–100% given the alignment score.

Superimposed filtration plots (Fig. 3) demonstrate that if the analyzed gene is close to the target protein, the weak filter provides better recognition. However, as the distance between the analyzed gene and the target increases, moderate filtration becomes beneficial. An explanation for this phenomenon is that stronger filtration decreases the number of candidate exons and thus eliminates competitors for the true exons when the similarity is low. However, the strong filtration (one candidate exon per 100 nucleotides) loses too many true exons, and its performance is inferior compared both to the weak and to the moderate filtration at the entire range of distances (data are not shown).

The same results can be seen on Table 7, in which results of predictions with the alignment score higher than 30% are given. These data confirm the above observation: weak filtration provides better results with mammalian targets, whereas moderate filtration is preferable with more distant target groups.

DISCUSSION

An important feature of the spliced alignment algorithm is the possibility of estimating the reliability of an

TABLE 6

Results of Prediction for Different Groups of Targets

Target	*N*	W	M	S
Human	256	99.5	98.4	94.9
Mammals	252	96.9	96.4	93.2
Birds	94	88.2	89.7	88.0
Cold-blooded vertebrates	98	87.9	88.3	87.2
Invertebrates	60	76.8	76.7	71.6
Other animals	39	78.0	78.3	73.1
Plants	37	86.2	86.9	82.6
Fungi	45	84.9	85.4	78.6
Other eukaryotes	14	84.1	85.5	75.2
Prokaryotes	32	79.9	81.7	78.3

Note. The average correlation coefficients are shown. Columns: *N*—number of genes with targets from the given group; W—weak filtration ($F = 14$, $E = 2$, partition function rescoring); M—moderate filtration ($F = 33$, $E = 3$, best structure rescoring); S—strong filtration ($F = 100$, $E = 3$, best structure rescoring). The first lane ("human") corresponds to the spliced alignment with the encoded protein itself as the target and is presented to demonstrate the influence of overfiltration.

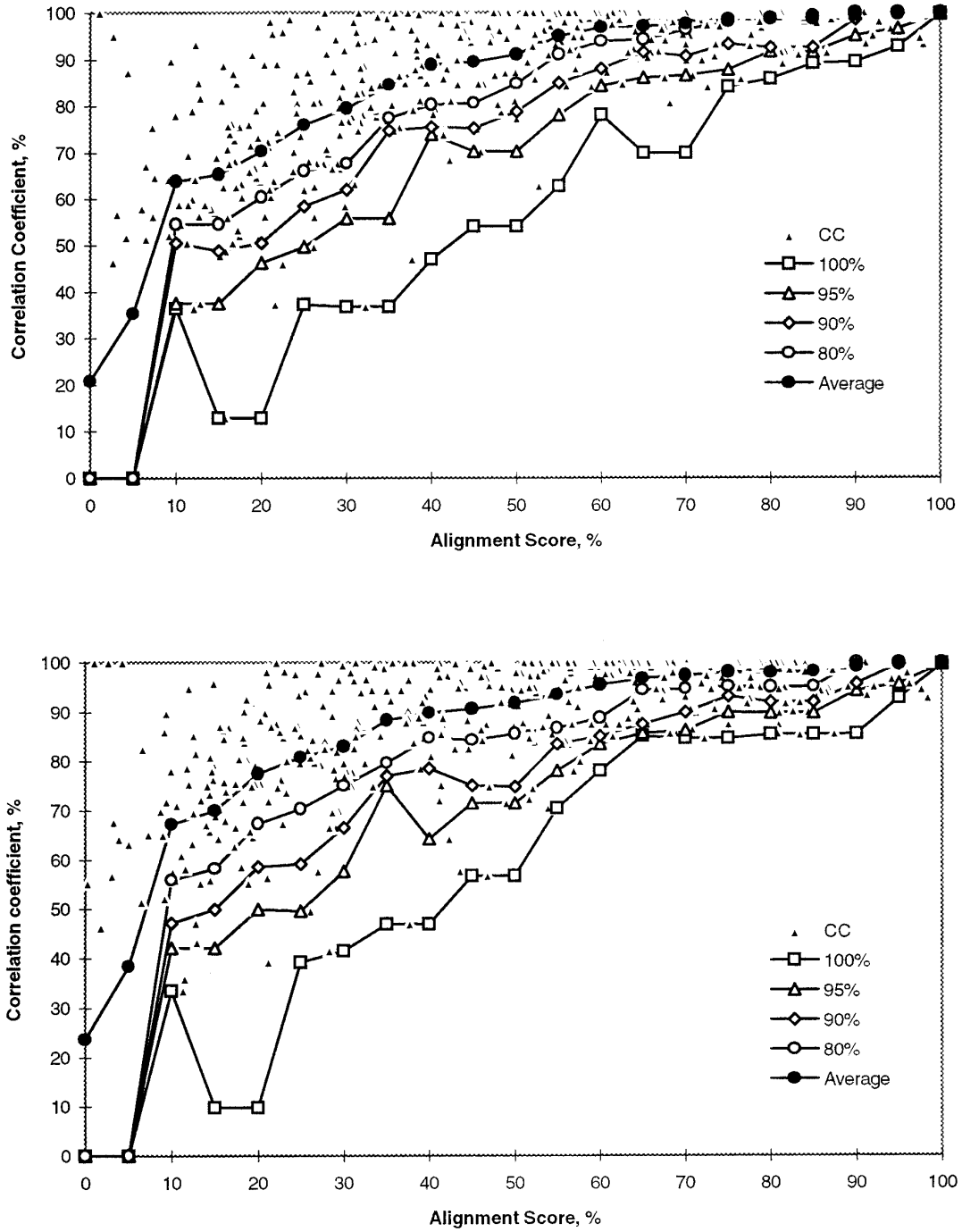

FIG. 1. Dependence of the correlation coefficient on the spliced alignment score. Scatter plot: spliced alignment score (horizontal axis), correlation coefficient (vertical axis); numerous points in the right upper corner are suppressed. Upper curve: average correlation coefficient. Other curves: correlation coefficient guaranteed with certainty $p = 100, 95, 90, 80$ (upward). If a curve p passes a point (s, c), then at least p among predictions with the score s have the correlation coefficient exceeding c. Top plot: weak filtering. Bottom plot: moderate filtering.

individual prediction by the spliced alignment score. Plot of the alignment score along the sequence (provided by Procrustes WWW server) allows one to view the relatively more or less reliable regions of the prediction. A different approach to the estimation of prediction reliability, based on construction of suboptimal spliced alignments and assigning the quality of fit to individual exons, is described in Sze and Pevzner (1997).

Most errors of the spliced alignment occur when there are unrelated domains in the target and the analyzed gene. This situation can be diagnosed by a very low spliced alignment score, and indeed, comparison of Tables 6 and 7 demonstrates that setting a recognition threshold sharply improves the average correlation between predicted and actual genes. In such cases it is reasonable to perform local spliced

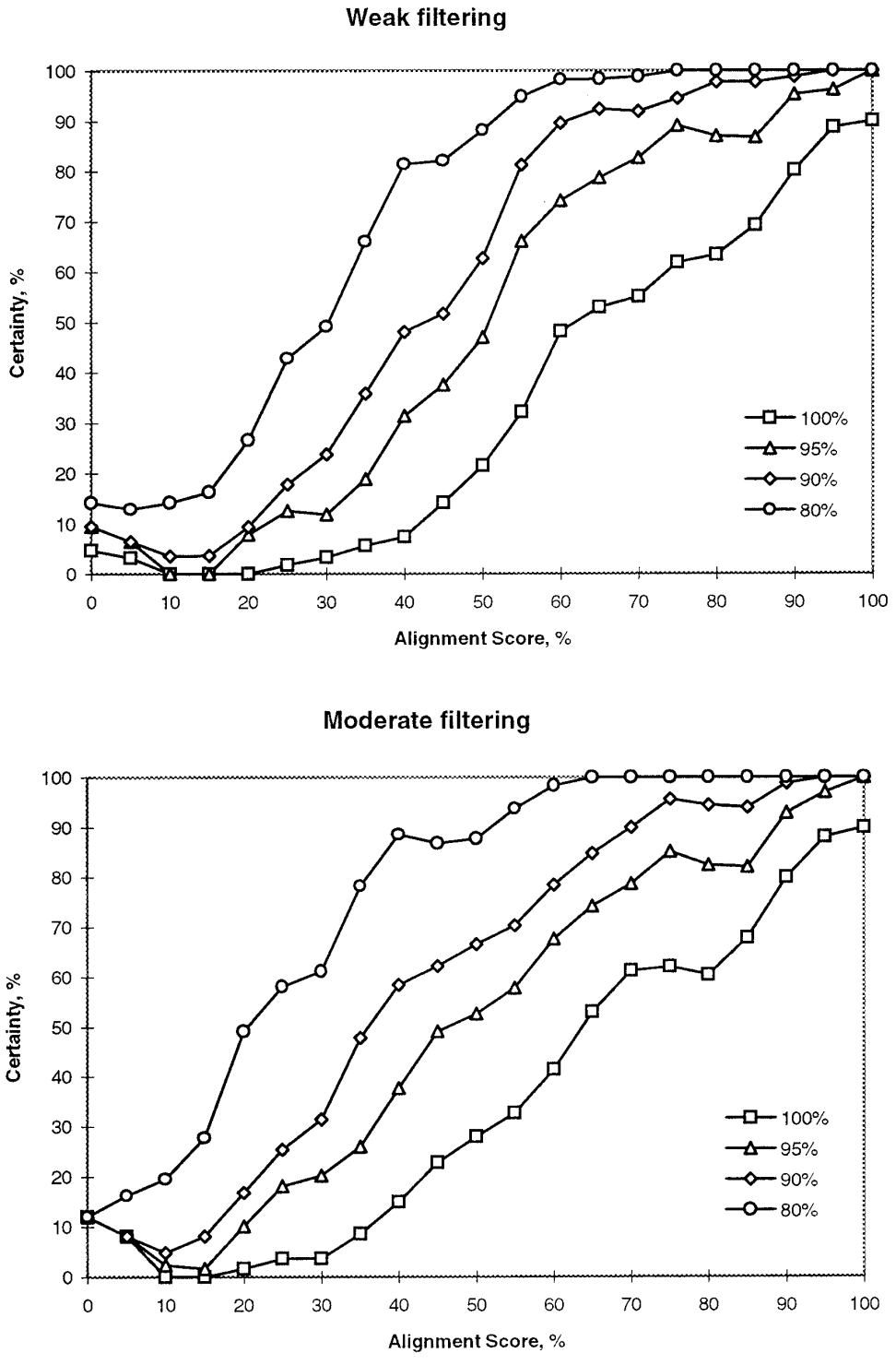

FIG. 2. Certainty level for predictions given the spliced alignment score. Horizontal axis—alignment score. Vertical axis—proportion of sequences with the correlation coefficient exceeding the given threshold. Plots correspond to thresholds c = 80, 90, 95, 100% (from top down). If a curve c passes a point (s, p), then at least p among predictions with the score s have the correlation coefficient exceeding c. Top plot: weak filtering. Bottom plot: moderate filtering.

alignment of the conserved regions only. This is an objective for further development.

Based on computer simulations, Gelfand *et al.* (1996a) suggested that spliced alignment with relatively strong filtration will perform better for distant targets, despite the risk of losing some true exons due to overfiltration. This conjecture was based on the following reasoning: the loss of some true exons is justified by the strong reduction of the number of candidate exons with the consequent decrease of combinatorial

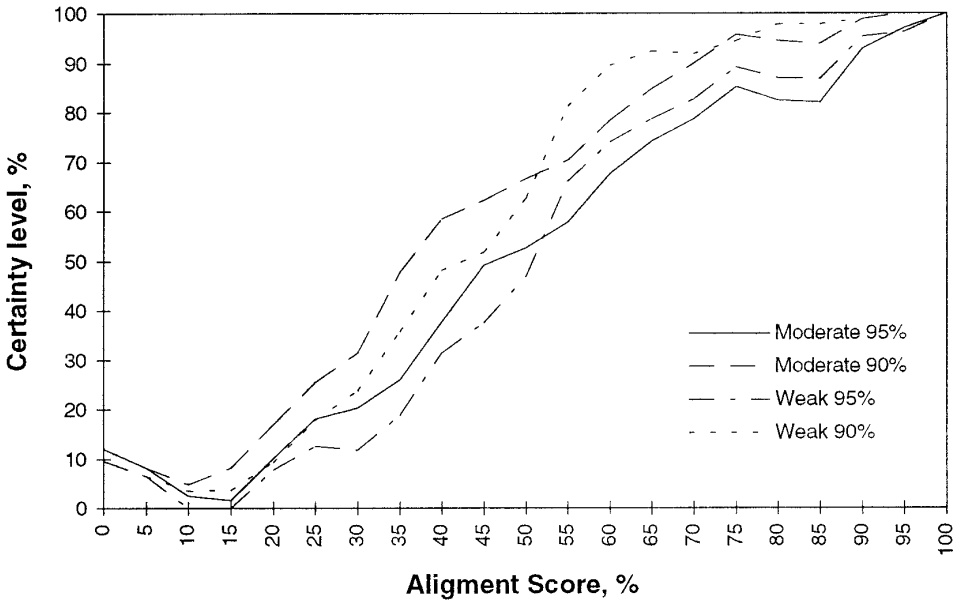

FIG. 3. Certainty level for predictions given the spliced alignment score for different filtration modes. Axes as for Fig. 2.

number of variants for the spliced alignment algorithm and sharp improvement of the alignment-based predictions for diverged targets.

This conjecture has been confirmed by the present analysis which used not only close, but some very distant targets (Table 1). However, excessively strong filtration (one candidate exon per 100 nucleotides) loses too many true exons. The tradeoff between overfiltration and excessive combinatorial flexibility depends on the similarity level, and the shift from weak to moderate filtration occurs at approximately 50% similarity level, as measured by the spliced alignment score (Fig. 3). It should be noted, however, that even for very close targets some filtration is necessary due to the *mosaic effect* (Sze and Pevzner, 1997).

Further development of Procrustes is directed toward construction of the local spliced alignment for analysis of genomic fragments containing incomplete genes (this is simple algorithmically, but additional work is required to derive scoring schemes and reliability indicators) and spliced alignment with nucleic acid targets applicable for gene recognition given noisy EST data.

SYNOPSIS

The current version of Procrustes http://www-hto.usc.edu/software/procrustes analyzes complete genomic and target sequences. For close targets the weak (default) filtering should be used and Fig. 2 can be used to estimate the reliability of predictions. If a distant target is used (spliced alignment score is below 50%) the spliced alignment should be repeated with moderate filtration (one exon per 33 nucleotides). Predictions with a score less than 30% should be considered tentative. Whenever possible, spliced alignment should be done with several targets. The spliced alignment plots (available through Procrustes WWW server) provide additional information about the prediction quality. In particular, a sharp local drop of the score for close targets is an indication of exon loss or substitution by

TABLE 7

Results of Prediction for Different Groups of Targets with 30% Alignment Score Threshold

Target	*N* (W)	W	*N* (M)	M	*N* (S)	S
Human	256	99.5	256	98.4	256	94.9
Mammals	243	98.0	243	97.7	239	94.6
Birds	70	96.4	68	96.5	65	93.9
Cold-blooded vertebrates	72	92.8	71	93.4	67	92.6
Insects	38	94.1	36	94.3	30	92.1
Other animals	26	93.2	25	96.5	23	92.2
Plants	24	95.2	24	94.2	21	89.4
Fungi	31	93.2	30	92.5	28	88.4
Other eukaryotes	10	96.1	10	96.6	8	90.6
Prokaryotes	15	87.4	14	90.5	10	88.0

Note. The average correlation coefficients are shown. Columns: *N*—number of genes with targets from the given group is shown for each mode of filtration (columns 2, 4, and 6); the remaining notation is as in Table 6.

spurious exons due to overfiltering. In this case, the weakest possible filtration (one exon per 10 nucleotides) can be attempted.

ACKNOWLEDGMENTS

We are grateful to Paul Hardy, Sergei Rahmanov, and Sing-Hoi Sze for many useful discussions and to Tatiana Astakhova for assistance in compiling the test sample. This work was supported by the U.S. Department of Energy under Grant DE-FG02-ER61919, Russian State Scientific Program "Human Genome," and Russian Fund of Basic Research.

REFERENCES

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219:** 555–565.

Benson, D. A., Boguski, M. S., Lipman, D. J., and Ostell, J. (1997). GenBank. *Nucleic Acids Res.* **25:** 1–6.

Burset, M., and Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34:** 353–367.

Fickett, J. W. (1996a). The gene identification problem: An overview for developers. *Comput. Chem.* **20:** 103–118.

Fickett, J. W. (1996b). Finding genes by computer: The state of the art. *Trends Genet.* **12:** 316–320.

Gelfand, M. S. (1995). Prediction of function in DNA sequence analysis. *J. Comput. Biol.* **2:** 87–115.

Gelfand, M. S., Mironov, A. A., and Pevzner, P. A. (1996a). Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA* **93:** 9061–9066.

Gelfand, M. S., Podolsky, L. I., Astakhova, T. V., and Roytberg, M. A. (1996b). Recognition of gene in human DNA sequences. *J. Comput. Biol.* **3:** 223–234.

Gish, W., and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272.

Hultner, M., Smith, D. W., and Wills, C. (1994). Similarity landscapes: A way to detect many structural and sequence motifs in both introns and exons. *J. Mol. Evol.* **38:** 188–203.

Jackson, I. J. (1991). A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res.* **19:** 3795–3798.

Rogozin, I. B., Milanesi, L., and Kolchanov, N. A. (1996). Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci.* **12:** 161–170.

Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996). *Entrez:* Molecular biology database and retrieval system. *Methods Enzymol.* **266:** 141–162.

Sharp, P. A. (1994). Split genes and RNA splicing. *Cell* **77:** 805–815.

Snyder, E. E., and Stormo, G. D. (1995). Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* **248:** 1–18.

Stoesser, G., Sterk, P., Tuli, M. A., Stoehr, P., and Cameron, G. N. (1997). The EMBL nucleotide sequence database. *Nucleic Acids Res.* **25:** 7–13.

Sze, S.-H., and Pevzner, P. A. (1997). Las Vegas algorithms for gene recognition: Subotimal and error tolerant spliced alignment. *J. Comput. Biol.* **4:** 297–310.

Uberbacher, E. C., and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor–neural network approach. *Proc. Natl. Acad. Sci. USA* **88:** 11261–11265.

Uberbacher, E. D., Xu, Y., and Mural, R. J. (1996). Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol.* **266:** 259–281.