BIOINFORMATICS

Vol. 17 no. 11 2001
Pages 1011–1018

Gene recognition in eukaryotic DNA by comparison of genomic sequences

P. S. Novichkov, M. S. Gelfand and A. A. Mironov*

State Scientific Center GosNIIGenetika, 1 Dorozhny pr. 1, Moscow 113545, Russia

Received on December 18, 2000; revised on April 9, 2001; accepted on July 25, 2001

ABSTRACT

Motivation: Sequencing of complete eukaryotic genomes and large syntenic fragments of genomes makes it possible to apply genomic comparison for gene recognition.

Results: This paper describes a spliced alignment algorithm that aligns candidate exon chains of two homologous genomic sequence fragments from different species. The algorithm is implemented in Pro-Gen software. Unlike other algorithms, Pro-Gen does not assume conservation of the exon–intron structure. Amino acid sequences obtained by the formal translation of candidate exons are aligned instead of nucleotide sequences, which allows for distant comparisons. The algorithm was tested on a sample of human–mammal (mouse), human–vertebrate (*Xenopus*) and human–invertebrate (*Drosophila*) gene pairs. Surprisingly, the best results, 97–98% correlation between the actual and predicted genes, were obtained for more distant comparisons, whereas the correlation on the human–mouse sample was only 93%. The latter value increases to 95% if conservation of the exon–intron structure is assumed. This is caused by a large amount of sequence conservation in non-coding regions of the human and mouse genes probably due to regulatory elements.

Availaility: Pro-Gen v. 3.0 is available to academic researchers free of charge at http://www.anchorgen.com/pro_gen/pro_gen.html.

Contact: misha@imb.imb.ac.ru

INTRODUCTION

Three main approaches to gene recognition exist. Statistical algorithms use local (splicing signals, promoters, polyadenylation sites) and global (e.g. codon usage) compositional features to find the optimal parse of a genomic fragment into protein-coding and non-coding regions (Gelfand, 1995; Burset and Guigo, 1996; Burge and Karlin, 1998). Similarity-based methods find regions similar (after translation) to already known proteins (Gish and States, 1993; Gelfand *et al.*, 1996; Birney and Durbin, 1997; Laub and Smith, 1998; Mironov *et al.*, 1998; Pachter *et al.*, 1999; Gotoh, 2000; Usuka and Brendel, 2000). The third approach is to align genomic sequences to ESTs in order to infer the exon–intron structure (Mott, 1997; Florea *et al.*, 1998; Jiang and Jacob, 1998; Mironov *et al.*, 1999; Usuka *et al.*, 2000).

These methods are useful, but each of them has its limitations. Although the specificity and sensitivity of individual predictions by statistical methods has increased since the benchmark study (Burset and Guigo, 1996), especially due to the appearance of hidden Markov model algorithms such as GenScan (Burge and Karlin, 1997), their reliability still is insufficient, especially in the context of genome projects generating long multigene fragments (Burge and Karlin, 1998). Similarity-based methods are the most reliable if a sufficiently close protein is available (Mironov *et al.*, 1998), but they cannot be used for genes encoding new proteins. EST-based algorithms have problems with artifactual ESTs (Tsai *et al.*, 1994; Wolfsberg and Landsman, 1997; Bouck *et al.*, 1999) and besides they cannot help in analysis of genes not represented in clone libraries because of narrow stage and tissue specificity of these genes.

Sequencing of large fragments of genomic DNA, and even complete eukaryotic chromosomes, makes it possible to apply comparison of genomic sequences for identification of protein-coding regions. This approach is based on the fact that protein-coding regions evolve much slower than non-coding regions. Thus candidate exons are seen as islands of similarity in alignment of genomic sequences harboring homologous genes. In particular, this approach has been successfully used to find new genes in syntenic regions of the human and mouse genomes (Ansari-Lari *et al.*, 1998) and in the genomes of nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* (Thacker *et al.*, 1999; Kent and Zahler, 2000). However, both studies relied on visual inspection of nucleotide sequence alignments (Ansari-Lari *et al.*, 1998) or dot plots (Thacker *et al.*, 1999).

Recently several algorithms have been described for automated gene recognition by genomic comparison. The ROSETTA algorithm (Batzoglou *et al.*, 2000a,b) first performs alignment of nucleotide sequences using the

*To whom correspondence should be addressed.

GLASS system, then identifies candidate exons (more exactly, exon pairs) within conserved regions, and then finds the optimal pair of candidate exon chains using the dynamic programming to maximize the score dependent on a set of statistical and similarity parameters.

The Conserved Exon Method (CEM; Bafna and Huson, 2000) identifies the candidate exon pairs using alignment of amino acid sequences derived by formal translation of genomic sequences. The exon–intron structure is constructed by dynamic programming as a chain of conserved exons maximizing the alignment score.

Thus both ROSETTA and CEM start with identification of conserved exon pairs. CEM seems to be more robust, since it relies on protein, and not nucleotide, alignment and thus it should be able to handle more distant comparisons. One more recently announced method, SGP (Wiehe *et al.*, 2000), starts with nucleotide alignment and then restricts it to a chain of segments bounded by potential splicing sites in both sequences. All these methods imply conservation of exon–intron structure and thus are intended for analysis of human–mouse (or, more generally, vertebrate) gene pairs.

We have developed an algorithm for automated gene recognition by comparison of genomic sequences, implemented in program Pro-Gen. In each sequence, it finds a chain of exons most similar to each other on the protein level. Unlike other programs, Pro-Gen does not assume conservation of the exon–intron structure and thus can be applied to analysis of relatively distant homologs, provided the amino acid sequence is conserved. In a preliminary study of simulated sequences Pro-Gen was shown to successfully predict genes at the 50% identity level with different exon–intron structure (Novichkov *et al.*, 2000a). Similar ideas in a simpler situation were implemented in the spliced alignment algorithm (Gelfand *et al.*, 1996). We present results of Pro-Gen testing on close (human–mouse), intermediate (human–*Xenopus*), and distant (human–*Drosophila*) gene pairs. In the human–mouse case, the results with and without assumption of conservation of the exon–intron structure are presented. All results are also compared with the theoretical upper limit as estimated by assuming one of the genes to be known.

Preliminary results of this study were reported in Novichkov *et al.* (2000b).

ALGORITHM

Pro-Gen accepts as input two genomic sequences S and T containing homologous genes. Initial identification of such gene pairs can be done using already existing programs such as TBLASTX (Altschul *et al.*, 1994).

The aim is to find the optimal alignment of exon chains (*spliced alignment*). For simplicity the exposition below skips technical details such as accounting for different reading frames. These details are trivial, although cumbersome. We also follow the computational biology tradition of using the term 'exon' to mean 'translated part of exon'. All alignments are performed over amino acid sequences obtained by formal translation, whereas signals (start and stop codons and candidate splicing sites) are analyzed on the nucleotide level. However, we will allow ourselves some liberty of speech using phrases like 'alignment ending at stop codon' when it does not create confusion. Similarly, 'matching of codons' means 'matching of amino acids encoded by these codons'.

At the first step, start and stop codon positions are determined and candidate donor and acceptor sites are predicted in both sequences. In the present version, candidate sites are defined using a positional weight matrix, although a more sophisticated procedure based on site statistics is possible (cf. Usuka and Brendel, 2000). The optimal spliced alignment is constructed using dynamic programming (Roytberg *et al.*, 1997). The main recurrency for computing the spliced alignment score σ is (Figure 1)

$$\sigma(m,n) = \max \begin{cases} \sigma(m-3,n-3) + \Phi(s_m,t_n) & \text{(matching codons } [s_m s_{m+1} s_{m+2}] \text{ and } [t_n t_{n+1} t_{n+2}]) \quad \text{(a)} \\ \sigma(m,n-3) - \Gamma & \text{(deletion in } S \text{ against codon } [t_n t_{n+1} t_{n+2}] \text{ in } T) \quad \text{(b)} \\ \sigma(m-3,n) - \Gamma & \text{(deletion in } T \text{ against codon } [s_m s_{m+1} s_{m+2}] \text{ in } S) \quad \text{(c)} \\ \mu_m(n-3) - \Delta & \text{if } m \text{ is an acceptor position (intron in } S) \quad \text{(d)} \\ \mu_m(n) - \Delta - \Gamma & \text{if } m \text{ is an acceptor position (intron in } S\text{, deletion in } T) \quad \text{(e)} \\ \nu_n(m-3) - \Delta & \text{if } n \text{ is an acceptor position (intron in } T) \quad \text{(f)} \\ \nu_n(m) - \Delta - \Gamma & \text{if } n \text{ is an acceptor position (intron in } T\text{, deletion in } S) \quad \text{(g)} \\ \lambda_{mn} - 2\Delta & \text{if } m, n \text{ are acceptor positions (introns in } S \text{ and } T) \quad \text{(h)} \end{cases} \tag{1}$$

where $\Gamma = \Gamma_{\text{gap}}$ is the penalty for the first deleted symbol (gap initialization) and $\Gamma = \Gamma_{\text{del}}$ is the penalty for gap extension; Δ is the intron penalty; $\Phi(A, B)$ is the weight for matching codons A and B. The variables μ, ν and λ keep the scores of best current alignments ending at donor sites respectively in S, T, or both sequences:

$$\begin{aligned} \mu_m(n) &= \max\{\sigma(i,n) | i < m, i \text{ is a donor position in } S\}, \quad \text{(a)} \\ \nu_n(m) &= \max\{\sigma(m,j) | j < n, j \text{ is a donor position in } T\}, \quad \text{(b)} \\ \lambda_{mn} &= \max\{\sigma(i,j) | i < m, i \text{ is a donor position in } S, \\ &\qquad j < n, j \text{ is a donor position in } T\}. \quad \text{(c)} \end{aligned} \tag{2}$$

The number of candidate sites is linear with respect to the sequence length. Thus the straightforward implementation of the above recurrency needs a quadratic number of comparisons in each cell. The number of cells is the

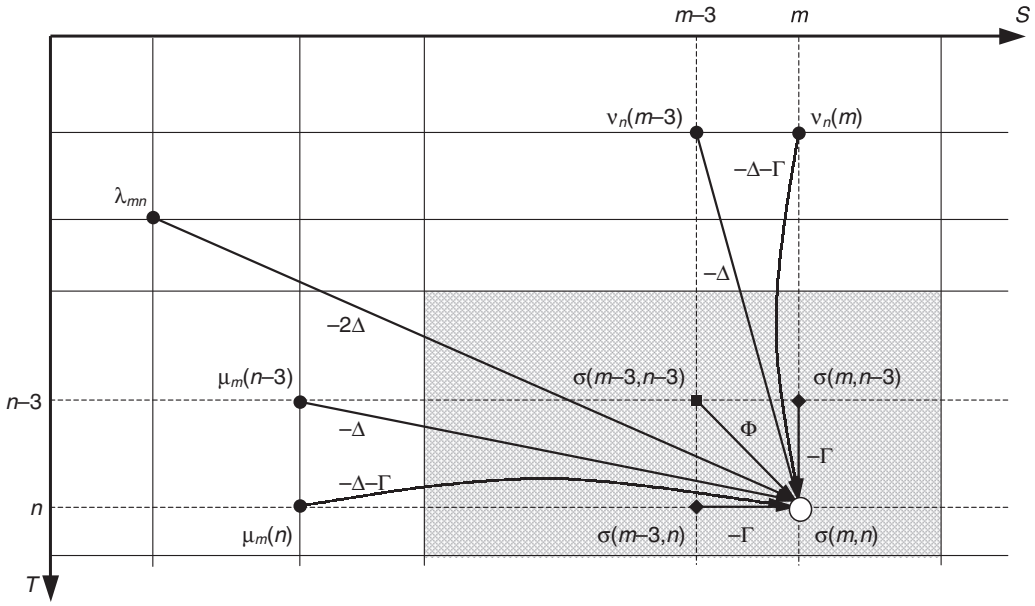

Fig. 1. The dynamic programming graph. Axes: positions in sequences S and T. All weighted arcs leading to the vertex (m, n) are shown. Shaded area: region of constancy of λ, μ and $\nu \cdot \sigma$: vertex scores. Other notation as in text.

product of the sequence lengths and thus the total time is proportional to the fourth power of the sequence length. However, note that the variables μ, ν and λ need to be re-computed only once each time the current position is a candidate donor site, since it is clear that these values are constant within rectangles bounded by candidate donor sites (the rectangle that covers the current position is shaded at Figure 1). Thus the computation time is proportional to the product of the sequence lengths.

Start and stop codons are treated in a specific manner. Paired ATG codons not only can be aligned as methionines, but may initialize an alignment. In-frame stop codons necessarily terminate exon chains. Any chain of exons must begin at a start codon and end at a stop codon. Split codons at exon boundaries are ignored.

To implement restrictions on the intron length, it is sufficient to modify the subscripts of μ, ν and λ in formula (1), using $\mu_{m-\mathrm{MIL}}$, $\nu_{n-\mathrm{MIL}}$ and $\lambda_{m-\mathrm{MIL},\, n-\mathrm{MIL}}$ in lines (d–h), where MIL is the minimum intron length. To take into account the similar restriction on the exon length, one can require in formula (2) that the chains end at exons exceeding the minimum allowed exon length. In the present version this is implemented by forcing the alignment to be extended if the last exon is too short. This is a non-optimal solution that may cause problems with weakly similar exons containing conserved cores. However, in practice this solution proved to be sufficient.

In the tests, the following values of the parameters were used: $\Gamma_{\mathrm{gap}} = 11$, $\Gamma_{\mathrm{del}} = 4$, $\Delta = 6$, $\Phi = \mathrm{PAM120}-1$ (each element of the standard PAM120 matrix is decreased by 1), the minimum intron length is 70 for vertebrates and 55 for *Drosophila*, the minimum exon length is 25. These parameters and the positional weight matrices for the recognition of splicing sites were taken from Mironov *et al.* (1998). The thresholds were set so as not to lose any true sites (thus roughly 50% of GT and AG sites are retained). Eliminating the threshold leads to increase of the run time and decrease of specificity of predictions due to additional false exons in conserved non-coding regions (data not shown). Note that the intron penalty is less than the gap penalty, although the double intron penalty exceeds the gap penalty. This balance of penalties ensures against addition of spurious tiny exons, although even modest similarity between correct exons is sufficient to incorporate them into the predictions. Short initial and terminal exons are not affected by the intron penalty, since only complete structures are considered, and thus the presence of such exons is forced.

As expected from the theoretical analysis, the run time depends linearly on the product of sequence lengths (data not shown). The average time for comparison of two 10 kb sequences, each containing one gene, is 7 min on Pentium II/400 MHz.

DATA

Pro-Gen was tested on a sample of human–mouse, human–*Xenopus* and human–*Drosophila* gene pairs. The vertebrate gene pairs were taken from the HOVERGEN

Table 1. Data sets. (a) Number of exons in human–*Drosophila*, human–*Xenopus* and human–mouse gene pairs. Rows: number of exons in human genes. Rows: number of pairs with the given number of exons (the number of exons in the human–*Xenopus* and human–mouse samples is always the same)

Human	*Drosophila*						Human–*Xenopus*	Human–mouse
	2	3	4	5	6	17		
2	–	–	–	–	1	–	4	12
3	1	–	–	–	–	–	3	14
4	4	3	1	–	–	–	2	13
5	1	–	2	–	–	–	1	8
6	3	–	1	1	1	–	3	10
7	1	–	1	–	–	–	–	4
8	–	–	–	2	–	–	1	4
10	–	–	–	–	–	–	–	2
11	–	–	–	–	–	–	–	1
13	–	–	–	1	–	–	–	–
14	–	–	–	–	–	–	–	1
28	–	–	–	–	–	–	–	1
30	–	–	1	–	–	–	–	–
38	–	–	–	–	–	1	–	–

(b) Histogram of the protein lengths (the average length of the proteins in a pair is considered)

From	101	201	301	401	501	601	801	1001	1201	1501	2001	3001
To	200	300	400	500	600	800	1000	1200	1500	2000	3000	4000
#	9	12	11	12	13	15	8	14	7	4	2	1

(c) Histogram of the genomic fragment lengths

From	900	2001	3001	4001	5001	6001	8 001	10 001	15 001	20 001
To	2000	3000	4000	5000	6000	8000	10 000	15 000	20 000	23 000
Human	4	19	26	16	18	10	5	7	1	2
Others	13	26	18	18	11	12	2	6	1	1

database (Duret *et al.*, 1994). The human–*Drosophila* gene pairs were taken from the Berkeley *Drosophila* Project sample (Reese *et al.*, 2000). The non-coding margins were trimmed not to exceed 1000 nucleotides at both ends.

Only protein pairs with identity exceeding 50% were considered. This is motivated by the observation that lower identity usually corresponds to homologous domains rather than orthologous genes (all known pairs of human–mouse orthologs are more than 50% identical; Makalowski and Boguski, 1998) and the fact that in a pilot experiment with simulated sequences we have observed that performance of the algorithm worsens sharply if the identity of the proteins is less than 50% (Novichkov *et al.*, 2000a,b).

Additionally, we have observed that the protein lengths in pairs with lower identity were very different. This provides additional motivation to the above criterion, since the domain organization of orthologous genes of higher eukaryotes is usually conserved (Mushegian *et al.*, 1997).

The characteristics of the samples (number of exons, protein and DNA fragment lengths) are given in Table 1. In particular, Table 1 provides information about the number of exons in the human–*Drosophila* pairs.

RESULTS

We denote by TP the number of correctly predicted coding nucleotides; TN, the number of correctly predicted non-coding nucleotides; FP, the number of non-coding nucleotides predicted to be coding; FN, the number of missed coding nucleotides. Specificity $\mathrm{Sp} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ is the fraction of true coding nucleotides among all predicted coding nucleotides. Sensitivity $\mathrm{Sn} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ is the fraction of correctly predicted coding nucleotides among all coding nucleotides. The correlation between the predicted and correct structures is defined as

$$\mathrm{CC} = (\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN})/[(\mathrm{TP}+\mathrm{FP})\cdot(\mathrm{FN}+\mathrm{TN})\cdot(\mathrm{TP}+\mathrm{FN})\cdot(\mathrm{FP}+\mathrm{TN})]^{1/2}.$$

For good predictions Sp, Sn, and CC are close to 1 (Burset and Guigo, 1996).

Table 2. Testing results. Human–mouse, human–*Xenopus* and human–*Drosophila* gene pairs are considered. The quality is ascribed according to the worst of the two predictions (e.g. if the specificity of the predicted mouse gene is 85%, sensitivity of the mouse gene is 100%, and both the specificity and the sensitivity of the predicted human gene is 95% the prediction is counted as 'fair')

	Drosophila	*Xenopus*	Mouse	Mouse*
Number of pairs	24	14	70	70
Exact predictions	13	6	24	32
Good predictions (Sp and Sn >90%)	9	7	29	27
Fair predictions (Sp and Sn >80%)	2	1	9	6
Bad predictions	0	0	8	5
Correlation coefficient (CC%)	98.1	96.8	93.6	95.4
Specificity (Sp%)	99.4	98.4	92.2	94.8
Sensitivity (Sn%)	98.0	96.9	97.9	98.0

*__Note:__ coincidence of the exon–intron structures is assumed.

Table 3. Prediction results on the exon level

Predicted gene	Target gene	Total true exons	Total predicted exons	Correct exons	Wrong exons	Missed exons	Overlapping exons
Human	*Drosophila*	196	202	181	7	2	13
Drosophila	Human	90	98	75	12	5	10
Human	*Xenopus*	56	61	46	7	2	8
Xenopus	Human	56	61	47	7	2	7
Human	Mouse	351	412	286	71	9	56
Mouse	Human	351	408	284	67	10	57
Human*	Mouse*	351	388	306	44	7	38
Mouse*	Human*	351	388	302	44	7	41

*__Note:__ coincidence of the exon–intron structures is assumed.

The scatterplots showing the dependence of the correlation coefficient on the identity level of considered genes are presented in Figure 2. The average values of CC, Sp and Sn and other general characteristics of the prediction quality are given in Table 2. The results of predictions on the exons level are given in Table 3. It is clear that the quality of predictions sharply depends on the evolutionary distance between the analyzed genes. Thus, most human–mouse comparisons lead to overpredictions, with the average correlation not much better than that of the best statistical methods (Burge and Karlin, 1997). On the other hand, more distant comparisons (human–*Xenopus* and human–*Drosophila*) lead to much better predictions.

In order to determine the upper boundary of the correlation coefficient, we have analyzed the same samples using Procrustes (Mironov *et al.*, 1998). The latter algorithm performs spliced alignment of a genomic sequence with a related (*target*) protein. We have predicted each gene of a pair using the protein encoded by the second gene as a target. The results are presented in Table 4. It is clear that in the human–*Xenopus* and human–*Drosophila* cases the predictions by Pro-Gen are almost as good as the predictions by Procrustes. Thus there is no considerable drop of reliability when two genomic sequences are used instead of a genomic sequence and a protein sequence, at least at the considered similarity level.

Table 4. Comparison of Pro-Gen and Procrustes. Notation as in Table 2

Predicted gene	Target gene or protein	Pro-Gen			Procrustes		
		CC	Sp	Sn	CC	Sp	Sn
Human	*Drosophila*	98.3	99.4	97.8	98.3	98.2	98.8
Drosophila	Human	97.9	99.3	98.2	97.0	97.9	97.9
Human	*Xenopus*	96.9	98.4	96.8	97.1	97.6	97.6
Xenopus	Human	96.8	98.4	96.9	97.5	98.4	96.7
Human	Mouse	93.6	92.2	98.0	98.8	99.2	98.9
Mouse	Human	93.5	92.2	97.8	98.6	98.7	98.8
Human*	Mouse*	95.6	94.8	98.2			
Mouse*	Human*	93.1	94.8	97.8	as above		

*__Note:__ coincidence of the exon–intron structures is assumed.

The main types of errors for the human–mouse sample are overpredictions caused by conservation of non-coding regions. This leads either to generation of spurious exons, especially upstream of the actual genes, or to extension of exons. We have applied the same analysis using fragments starting at the correct ATG and ending at the correct stop codon (although knowledge of the correct start and stop codons was not assumed). In this case the quality of predictions in the human–mouse pairs was much higher (data are not shown).

DISCUSSION

Most errors, especially at close comparison, seem to concentrate upstream of the actual genes. Although some of these cases may in fact be caused by alternative splicing (Mironov *et al.*, 1999; Hanke *et al.*, 1999), the majority of errors is probably due to conservation of non-coding regions (Duret *et al.*, 1994; Lipman, 1997). Indeed, more than half of 5′-untranslated regions are covered by regions conserved between human and mouse (Jareborg *et al.*, 1999). Although some conservation can be seen at larger evolutionary distance (Gellner and Brenner, 1999), it is confined to relatively short regions comprising the binding sites of transcription factors (Duret and Bucher, 1997), and thus does not lead to creation of spurious exons. Another source of errors is due to very short initial and terminal exons (that can be as short as three nucleotides). This leads both to overprediction (if stronger extensions are found) and underprediction (if neither the correct exons, nor positive-scoring alternatives can be found).

The results on the human–mouse data are worse than

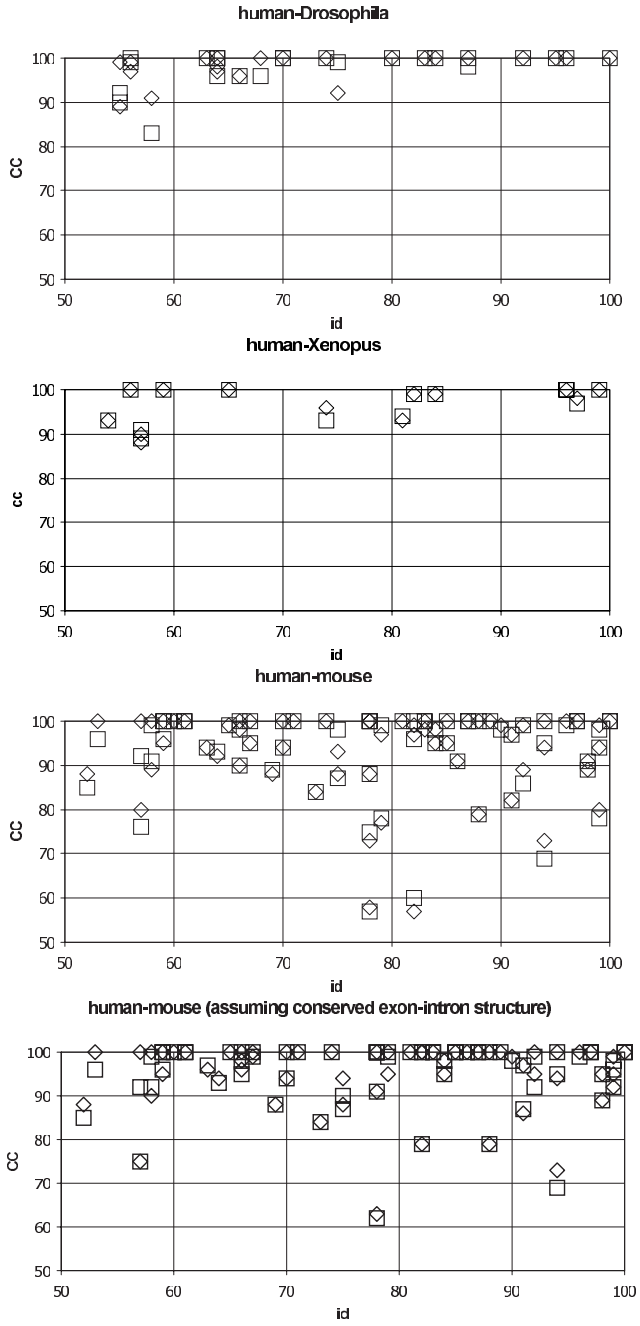

Fig. 2. Dependence of the prediction results on the level of identity of true proteins. Horizontal axis: identity (in %); vertical axis: correlation coefficient (in %). Squares: human genes. Diamonds: *Drosophila*, *Xenopus*, mouse genes.

those demonstrated by three recently published programs, ROSETTA (Batzoglou *et al.*, 2000a,b), CEM (Bafna and Huson, 2000), and SGP (Wiehe *et al.*, 2000). However, if conservation of the exon–intron structure is assumed, the results of Pro-Gen on this sample are very close to those demonstrated by the three programs above. ROSETTA and SGP are described as specifically intended for analysis of the human–mouse gene pairs, and applying them to the human–*Xenopus* and human–*Drosophila* samples would mean putting them to a predictable disadvantage. In particular, these programs assume conservation of the exon–intron structure, which is usually correct for comparison of the vertebrate genes, but not for more distant gene pairs. Table 1 demonstrates that the number of introns in the *Drosophila* genes from the human–*Drosophila* pairs is much lower than in the human genes (the total numer of introns is 66 and 172 respectively). Indeed, positions of only 44 introns coincide up to three codons, thus at most two thirds of the *Drosophila* introns and one quarter of the human introns are conserved between the two genomes. Anectodal evidence shows that the exon–intron structure may differ between orthologous insect genes (Novichkov *et al.*, 2000a), although we are not aware of any systematic studies in this area. Scattered as they are, in our opinion these observations are sufficient to warrant development of algorithms insensitive to changes in the exon–intron structure. CEM is more general than ROSETTA and SGP, but it does not seem to be available in the public domain. Thus we did not perform detailed comparison of Pro-Gen with these programs. On the other hand, the DNA–protein spliced alignment program Procrustes sets the upper threshold on the correlation coefficient. Table 4 demonstrates that the Pro-Gen performance is very close to this threshold.

One possible difficulty not addressed in this study, but important in practical applications, is analysis of multigene genomic fragments. In the human–mouse case, where the order of orthologous genes is conserved in syntenic regions, this can lead to creation of chimeric exon chains due to merging of adjacent genes. None of the current genome-comparison programs seem to address this problem directly. However, this should not be a problem for analysis of more distant genomes. Recent analyses demonstrate that although some synteny is conserved in all vertebrate comparisons (in particular, between human and fugu fish; Gellner and Brenner, 1999; Brunner *et al.*, 1999), it is not conserved in more distant pairs from various taxonomic groups. For instance, it has been reported that there are little collinear gene chains in rice and *Arabidopsis* (Devos *et al.*, 1999). There is no observable synteny between the human and *Drosophila* genomes. Even in the fugu–human case the syntenic gene chains are often interrupted by insertion of non-homologous genes (Gellner and Brenner, 1999).

Thus the optimal distance between the genomes for the Pro-Gen analyses is determined by a tradeoff between two requirements: the entire protein has to be conserved, as opposed to domain conservation, but the conservation of non-coding regions and the level of synteny should be low. Probably the best counterparts for analyses of the human

genome are cold-blooded vertebrates for fine mapping of the exon boundaries and invertebrates for identification of boundaries between genes. Of course, in all cases it makes sense to use as many pairs as possible for analysis of a single gene: coincidence of predictions in different pairs will be a strong identification of the prediction validity. If only a close (e.g. mouse) genome fragment is available for comparative analysis, it will make sense to use Pro-Gen in combination with a statistical algorithm that can eliminate spurious similarities in non-coding regions. Although the existing methods for prediction of functional sites, in particular, promoters of transcription, are not sufficiently reliable yet (Fickett and Hatzigeorgiou, 1997), a combination of a promoter recognition program and a similarity-based gene recognition program may prove more successful than independent application of these approaches.

ACKNOWLEDGEMENTS

We are grateful to Jim Fickett, Michael Roytberg, and Pavel Pevzner for useful discussions.

This study was partially supported by grants from the Russian State Scientific Program 'Human Genome', INTAS (99-1476), the Howard Hughes Medical Institute (55000309), and the Ludwig Institute for Cancer Research.

REFERENCES

Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genet.*, **6**, 119–129.

Ansari-Lari,M.A., Oeltjen,J.C., Schwartz,S., Zhang,Z., Muzny,D.M., Lu,J., Gorrell,J.H., Chinault,A.C., Belmont,J.W., Miller,W. and Gibbs,R.A. (1998) Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.*, **8**, 29–40.

Bafna,V. and Huson,D.H. (2000) The conserved exon method for gene finding. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology ISMB'2000*. pp. 3–12.

Batzoglou,S., Pachter,L., Mesirov,J., Berger,B. and Lander,E.S. (2000a) Human and mouse gene structure: comparative analysis and application to exon prediction. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology RECOMB'2000*. pp. 46–53.

Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B. and Lander,E.S. (2000b) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.

Birney,E. and Durbin,R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. In *Proceedings of the 5th Conference on Intelligent Systems in Molecular Biology ISMB'97*. pp. 56–64.

Bouck,J., Yu,W., Gibbs,R. and Worley,K. (1999) Comparison of gene indexing databases. *Trends Genet.*, **15**, 139–162.

Brunner,B., Todt,T., Lenzner,S., Stout,K., Schuulz,U., Ropers,H.-H. and Kalscheuer,V.M. (1999) Genomic structure and comparative analysis of nine *Fugu* genes: conservation of synteny with human chromosome Xp22.2-p22.1. *Genome Res.*, **9**, 437–448.

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

Burge,C.B. and Karlin,S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, **8**, 346–354.

Burset,M. and Guigo,R. (1996) Evaluation of gene recognition porgrams. *Genomics*, **34**, 353–367.

Devos,K.M., Beales,J., Nagamura,Y. and Sasaki,T. (1999) *Arabidopsis*-rice: will colinearity allow gene prediction across the eudicot-monocot divide? *Genome Res.*, **9**, 825–829.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.

Duret,L., Mouchiroud,D. and Gouy,M. (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.*, **22**, 2360–2365.

Fickett,J.W. and Hatzigeorgiou,A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.

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

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. (1996) Gene recognition via spliced alignment. *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.

Gellner,K. and Brenner,S. (1999) Analysis of 148 kb of genomic DNA around the wnt1 locus of Fugu rubripes. *Genome Res.*, **9**, 251–258.

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

Gotoh,O. (2000) Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps. *Bioinformatics*, **16**, 190–202.

Hanke,J., Brett,D., Zastrow,I., Aydin,A., Delbruck,S., Lehmann,G., Luft,F., Reich,J. and Bork,P. (1999) Alternative splicing of human genes: more the rule than the exception? *Trends Genet.*, **15**, 389–390.

Jareborg,N., Birney,E. and Durbin,R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.*, **9**, 815–824.

Jiang,J. and Jacob,H.J. (1998) EbEST: an automated tool using expressed sequence tags to delineate gene structure. *Genome Res.*, **8**, 268–275.

Kent,W. and Zahler,A. (2000) Conservation, regulation, synteny, and introns in a large-scale *C.briggsae–C.elegans* genomic alignment. *Genome Res.*, **10**, 1115–1125.

Laub,M.T. and Smith,D.W. (1998) Finding intron/exon splice junctions using INFO, INterruption Finder and Organizer. *J. Comput. Biol.*, **5**, 307–321.

Lipman,D.J. (1997) Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.*, **25**, 3580–3583.

Makalowski,W. and Boguski,M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA*, **95**, 9407–9412.

Mironov,A.A., Roytberg,M.A., Pevzner,P.A. and Gelfand,M.S. (1998) Performance-guarantee gene predictions via spliced alignment. *Genomics*, **51**, 332–339.

Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.

Mott,R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.*, **13**, 477–478.

Mushegian,A.R., Bassett,D.E. Jr., Boguski,M.S., Bork,P. and Koonin,E.V. (1997) Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl Acad. Sci. USA*, **94**, 5831–5836.

Novichkov,P.S., Gelfand,M.S. and Mironov,A.A. (2000a) Prediction of the exon–intron structure by comparison of genomic sequences. *Mol. Biol.*, **34**, 200–206.

Novichkov,P.S., Gelfand,M.S. and Mironov,A.A. (2000b) Pro-Gen: prediction of the exon–intron structure by comparison of genomic sequences. In *Proceedings of the 2nd International Conference on Bioinformatics of Genome Regulation and Structure BGRS'2000*, Vol. 2, pp. 42–43.

Pachter,L.S., Batzoglou,S., Spitkovsky,V.I., Banks,E., Lander,E.S., Kleitman,D.J. and Berger,B. (1999) A dictionary-based approach for gene annotation. *J. Comput. Biol.*, **6**, 419–430.

Reese,M.G., Hartzell,G., Harris,N.L., Ohler,U., Abril,J.F. and Lewis,S.E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, **10**, 483–501.

Roytberg,M.A., Astakhova,T.V. and Gelfand,M.S. (1997) Combinatorial approaches to gene recognition. *Comput. Chem.*, **21**, 229–235.

Thacker,C., Marra,M.A., Jones,A., Baillie,D.L. and Rose,A.M. (1999) Functional genomics in *Caenorhabditis elegans*: an approach involving comparisons of sequences from related nematodes. *Genome Res.*, **9**, 348–359.

Tsai,J.-Y., Namin-Gonzalez,M.L. and Silver,L.M. (1994) False association of human ESTs. *Nature Genet.*, **8**, 321–322.

Usuka,J. and Brendel,V. (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. *J. Mol. Biol.*, **297**, 1075–1085.

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

Wiehe,T., Gebauer-Jung,S., Abril,J. and Guigo,R. (2000) Comparative genomics: homology based gene identification and gene structure validation. In *Proceedings of the 2nd International Conference on Bioinformatics of Genome Regulation and Structure BGRS'2000*, Vol. 2, pp. 44–45.

Wolfsberg,T.G. and Landsman,D. (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.*, **25**, 1626–1632.