Biophysics, Vol. 48, Suppl. 1, 2003, pp. S71–S75.

GipsyGene: A Statistics-Based Gene Recognizer for Fungal Genomes

A. D. Neverov[1], M. S. Gelfand[1–3], and A. A. Mironov[1,3]

[1]State Scientific Center GosNIIGenetika, 1 1st Dorozhny proezd, Moscow, 117545 Russia; E-mail: neva_2000@mail.ru
[2]Institute for Problems of Information Transmission, Russian Academy of Sciences, 19 Bolshoi Karetny per., Moscow, 127994 Russia
[3]Department of Bioengineering and Bioinformatics, Moscow State University, 1-73 Vorobievy Gory, Moscow, 119992 Russia

Received December 7, 2003; in final form, January 9, 2004

Abstract—We developed a program GipsyGene for gene recognition in DNA of lower fungi. It is based on a Hidden Markov Model with duration states [1]. The program identifies genes on both DNA strands in one pass; at that, the DNA fragment may contain several genes, partial genes, or no genes at all. The program can be easily re-trained for new genomes. We developed a new statistical model of the branch point that considerable improved recognition of exon 5′ termini. The program was tested on genes of *Aspergillus* spp. and *Neurospora crassa*, and the reliability was shown to be comparable to that of GenScan on single gene human fragments. In addition, we report the results of testing on multigene fragments. The program is available upon request (neva_2000@mail.ru).

Key words: genome, fungi, gene recognition, Hidden Markov Models

INTRODUCTION

Lower fungi are important for agriculture, medicine, food industry, and ecology in general. Recently the genomes of several fungi, in particular, *Aspergillus nidulans*, *Neurospora crassa, Magnaporte grizea* have been sequenced [Fungal Genetics Stock Center] and it is clear that the sequencing would continue. Thus there arises a need in programs for identification of protein-coding genes in these genomes. Gene recognition programs can be divided into statistics-based ones and those based on similarity to known sequences, namely, proteins, mRNAs, and Expressed Sequence Tags. The latter usually produce more reliable results, but they cannot be applied to identification of new genes. As it is unlikely that EST collections will be available for many fungal species, and these organisms have a considerable number of taxon-specific proteins, statistical programs for gene identification should be developed. No such programs were available when this project was initiated.

Such programs can be divided into universal ones and programs requiring training on known genes. The reliability of the former is insufficient [2]. On the other hand, genes with known homologs can be used as a training set. Thus, our aim was to create a gene identification problem that could be easily trained for any given genome, but still would utilize some specifics of fungal genes.

ALGORITHM

The program is based on the Hidden Markov Chain with state durations, as in GenScan [1]. The model includes the following coding and non-coding states: single-exon gene; initial, middle, and terminal exon (more exactly, coding part of an exon); intron; intergenic region. The probability of a hidden state is computed as the probability of coding or non-coding state of the given sequence multiplied by the duration probability of the corresponding state.

The probabilities of donor and acceptor splicing sites are computed as in GenScan [1]. Depending on

the existence of significant correlations between the splicing signal positions, a statistical model is selected. If the correlations are absent or non-significant because of a small sample size, the probability of a site is calculated using the positional nucleotide frequency matrix. For acceptor sites, if most positions show significant correlations, the WAM model is used, in which the positional nucleotide probability depends on the previous nucleotide. For donor sites, if there are significant correlations between both adjacent and distanced positions, the probability is computed using the MDD model.

As for many fungal genomes it is difficult to compile a training set of sufficient size, we implemented several models for calculating the probability of coding and non-coding states. The program automatically selects the strongest applicable model. Coding DNA can be modeled using (1) codon statistics; (2) Markov chain of the first order for amino acids combined with statistics of synonymous codons; (3) three-periodical Markov chains of orders three through five. The models of non-coding DNA were the Markov chains of the first, third, and fifth order.

To improve recognition of intron–exon juctions (5′-termini of exons), we developed a model of the branch point. A candidate branch point is scored taking into account the profile model, distribution of the distances to the acceptor site, the presence and the number of AG dinucleotides between the branch point and the acceptor site. The profile and the distributions are computed basing on analysis of sites most similar to those of *Saccharomyces*. This procedure for constructing the profile was suggested in [3]. We modified it by increasing the number of positions taken into account (from 5 to 7) and adding pseudocounts. The latter correct the profile under assumption that some site variants are weakly represented in the training set because of the random sampling procedure.

In order to construct the branch point profile, the window (–40) to (–2) upstream of the acceptor site was scanned, and the site closest to the [CT]T[AG]a[CT] consensus (ideally CTAaC) was selected. Deviations in at most two out of three degenerate positions were allowed. The set of candidates was used to determine the distribution of branch point distances to the acceptor site. Application of the same procedure to random sequences sets the noise level, that is, the average frequency of the candidate sites in a random sequence. Then the window position was re-set so that the frequency of the candidate sites there exceeded the noise level. The set of candidates in the final word was used to construct the recognition profile of length 7 (two positions 5′ to the consensus were added), the distribution of distances between branch points and acceptor sites, and the number of AG dinucleotides in this interval. Each element of the obtained frequency matrix was augmented by the pseudocount equal to the square root of the sample size.

PROCEDURES

Construction of the Training and Testing Sets

The program was tested on single gene and multigene fragments of *Aspergillus* spp. and *Neurospora crassa* DNA. In a preliminary study, it was shown that genomic sequences of the *Aspergillus* species are statistically homogeneous. The training set consisted of 193 genes of *Emericella nidulans* (*Aspergillus nidulans*), whereas the testing set was formed by 252 single genes of various *Aspergillus* species. For *Neurospora crassa* the training and testing set consisted of 99 and 118 genes respectively. The testing and training sets did not intersect.

The procedure of constructing these sets was as follows. (1) All DNA sequences for the given organism were selected from GenBank. (2) Only sequence fragments containing a coding sequence were selected; sequences with in-frame stop codons or non-canonical splicing sites were discarded. (3) Sequences containing in the DEFINITION field keywords “contig,” “BAC,” “clone” were discarded. (4) Sequences with coinciding accession numbers in the “ACCESSION” field were removed. (5) The remaining sequences were divided into two sets of approximately equal size.

In order to test the program on multigene fragments, we generated a set of genome-like sequence fragments. They contained complete and partial genes of the *Aspergillus* testing set on both strands with probability 50%. Intergenic regions of average length approximately 2000 bp were imitated by concatenation of non-coding DNA segments. This procedure was necessary, as the available data do not include a sufficient number of well-annotated multigene fragments. The generated set contained true genes from GenBank and did not contain any missing genes, nor genes predicted by other programs, whereas the statistical properties of such fragments were the same as those of the genome in general.

Table 1. Results of GipsyGene testing on genes of *Aspergillus* spp. and *Neurospora crassa*

Aspergillus spp.									
Set	Nucleotide parameters					Exon parameters			
	CC	AC	Sn	Sp	QQ	Sn	Sp	MissE	WrongE
training	0.87	0.88	0.96	0.92	0.89	0.69	0.61	0.06	0.16
with BPM	0.89	0.89	0.96	0.92	0.9	0.81	0.72	0.04	0.14
testing	0.86	0.87	0.95	0.95	0.9	0.75	0.64	0.05	0.16
with BPM	0.9	0.9	0.96	0.95	0.92	0.8	0.73	0.04	0.12
genome-like	0.89	0.89	0.96	0.9	0.86	0.76	0.63	0.06	0.2
with BPM	0.89	0.9	0.96	0.9	0.87	0.81	0.69	0.05	0.18
Neurospora crassa									
Set	Nucleotide parameters					Exon parameters			
	CC	AC	Sn	Sp	QQ	Sn	Sp	MissE	WrongE
training	0.87	0.87	0.93	0.92	0.87	0.69	0.64	0.10	0.13
with BPM	0.88	0.88	0.94	0.93	0.88	0.72	0.64	0.07	0.14
testing	0.86	0.86	0.91	0.93	0.87	0.66	0.61	0.11	0.12
with BPM	0.89	0.89	0.93	0.95	0.89	0.75	0.69	0.08	0.12

Notes: BPM: branch point model. CC: correlation coefficient. AC: approximate correlation. Sn: sensitivity. Sp: specificity. MissE: missed exons. WrongE: false exons.

To assess the sensitivity of GipsyGene to genes in long DNA sequences, a different procedure was applied. Genes of the *Neurospora* testing set were identified in the complete genome [7] using the following procedure. Protein sequences obtained by translation of 118 coding regions were compared with the complete genome using TBLASTN [4] at the significance threshold EVAL < 10^{-30}. Thus identified 115 genome regions were aligned with the initial proteins using Pro-Frame [5], and the exon–intron structure of the genes was determined. The genome fragments obtained were used to determine the drop in sensitivity at long multigene fragments of *Neurospora crassa* as compared with single-gene fragments.

Assessing the Results of Prediction on Genes of *Neurospora crassa* Identified in the Genome DNA Sequence

Since the genome sequence could contain sequencing errors or belong to a different strain, Pro-Frame alignments could contain non-canonical sites at exon–intron junctions and hanging (not align) protein ends. Thus the sensitivity was computed only for canonical sites satisfying the GT-AG rule. For exons or introns with both canonical sites, the exon (resp. intron) sensitivity was computed as the fraction of complete segments of the given type with both sites correctly predicted. The fraction of missed exons (MissE) was computed relative to all exons identified by Pro-Frame.

Assessing the Quality of Predictions for Training, Testing, and Genome-like Sets

The quality of predictions at the nucleotide level (sensitivity Sn, specificity Sp, correlation coefficient CC, approximate correlation AC) and at the exon level (Sp, Sn, missed exons MissE, false exons WrongE) was calculated as in [1]. In addition, we considered the overlap coefficient QQ, the ratio of the overlap to the union of the true and predicted coding regions. The use of this measure is preferable for the analysis of single gene fragments, as it discounts the number of non-coding nucleotides. An exon was considered correctly identified if both its splicing sites were predicted correctly.

RESULTS

Prior to testing, optimal statistical models for the coding and non-coding models for *Aspergillus* genes were selected. Surprisingly, complicated models (up to the third-order Markov chain) were only slightly better than codon statistics (data not shown). The fifth-order Markov model is cenventionally considered to be preferable for distinguishing between exon and intron sequences. In our case, the training set is

Table 2. Nucleotide, exon, intron, and site sensitivity of GipsyGene on genomic fragments of *Neurospora crassa* containing genes from the testing set

Nucleotide sensitivity: SnN = 0.94
Exon sensitivity:
307 complete exons of 326 exons identified by Pro-Frame
Sensitivity SnE = 226/307 (0.74)
Missed MissE = 18/326 (0.06)
Intron sensitivity:
183 complete introns of 211 introns identified by Pro-Frame
Sensitivity SnI = 183/211 (0.87)
Site sensitivity:
Acceptor sites SnA = 187/213 (0.88)
Donor sites SnD = 187/212 (0.88)

Notes: Genes were identified by TBLASTN [4]; exons and introns, by Pro-Frame [5]. Exon and intron sensitivities were determined on complete exons and introns bounded by canonical GT and AG dinucleotides. Site sensitivity was computed relative to canonical sites identified by Pro-Frame.

too small, and, although this model outperformed all other models on the training set, its results on the testing set were weak. Thus the program was tested in the following configuration: three-periodical Markov model of the third order for coding regions, and homogeneous Markov model of the third order for noncoding DNA.

The performance parameters on the training and testing sets of *Aspergillus* and *Neurospora* are given in Table 1. An addition, this table contains the results of testing on the genome-like set of *Aspergillus*. In order to estimate the influence of the branch point model, the parameters both with and without this model are given.

One can see that the contribution of the branch point model is minor at the nucleotide level (about 1% QQ), but very considerable for exons: it increases the exon sensitivity on the *Aspergillus* and *Neurospora* testing sets of by 5% and 9% respectively, and the exon specificity by 9% and 8% respectively. Further we assume that the branch point model is included into the base configuration of the program, and all comparisons are done relative to the testing set in he base configuration. Testing of GipsyGene on the genome-like set of *Aspergillus* demonstrated a decrease of the nucleotide and exon specificity by 5% and 4% respectively compared with the testing set, and an increase of the false exon fraction by 6%. The sensitivity of the program did not change.

Table 2 contains sensitivity measures for nucleotides, exons, introns, donor and acceptor sites. The program was tested on long fragments of the *Neurospora crassa* genome containing genes from the testing set. Table 2 shows that GipsyGene does not lose the nucleotide and exon sensitivity on genomic sequences. The site sensitivity was 88% for sites of both types. The intron sensitivity (87%) is higher than the exon sensitivity (74%), maybe because introns are generally shorter than exons: the average lengths were 68 bp for introns and 187 bp for exons.

DISCUSSION

GipsyGene is a program for gene recognition in DNA of lower fungi. The program does not rely on data about homologs and thus can be used for recognition of new taxon-specific genes. The program can be trained on a relatively small sample of known genes or genes identified by similarity. As most lower fungi have a relatively well-defined branch point signal, the program incorporates a model of this signal, considerably increasing exon sensitivity. The quality of GipsyGene predictions of fungal genes is comparable to that of GenScan [1] on human genes. The program does not lose sensitivity on real genome

fragments. Like all programs of this class [6], GipsyGene loses specificity on multigene fragments, as currently there are no models for gene boundaries. Testing of the program on artificial multigene fragments with an average intergenic distance of 2000 bp (characteristic of *Aspergillus*) showed a decrease in specificity by about 5%.

ACKNOWLEDGMENTS

The authors are grateful to V.Yu. Makeev for useful discussions. The study was supported by grants from HHMI (55000309) and LICR (CRDF RB0-1268).

REFERENCES

1. Burge, C. and Karlin, S., Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol*., 1997, vol. 268, pp. 78–94.
2. Kraemer, E., Wang, J., Guo, J., Hopkins, S., and Arnold, J., An analysis of gene-finding programs for Neurospora crassa, *Bioinformatics*, 2001, vol. 17, pp. 901–912.
3. Clark, F., Thanaraj, T.A., Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human, *Hum. Mol. Genet*., 2002, vol. 11, pp. 451–464.
4. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search Programs, *Nucleic Acids Res*., 1997, vol. 25, pp. 3389–3402.
5. Mironov, A.A., Novichkov, P.S., and Gelfand, M.S., Pro-Frame: similarity-based gene recognition in eukaryotic DNA sequences with errors, *Bioinformatics*, 2001, vol. 17, pp. 13–15.
6. Guigo, R., Agarwal, P., *et al*., An assessment of gene prediction accuracy in large DNA sequences, *Genome Res*., 2000, vol. 10, pp. 1631–1642.
7. James E. Galagan, Sarah E. Calvo, Katherine A. Borkovich, Eric U. Selker, Nick D. Read, David Jaffe, William Fitzhugh, Li-Jun Ma, Serge Smirnov, Seth Purcell, Bushra Rehman, Timothy Elkins, Reinhard Engels, Shunguang Wang, Cydney B. Nielsen, Jonathan Butler, Matthew Endrizzi, Dayong Qui, Peter Ianakiev, Deborah Bell-Pedersen, Mary Anne Nelson, Margaret Werner-Washburne, Claude P. Selitrennikoff, John A. Kinsey, Edward L. Braun, Alex Zelter, Ulrich Schulte, Gregory O. Kothe, Gregory Jedd, Werner Mewes, Chuck Staben, Edward Marcotte, David Greenberg, Alice Roy, Karen Foley, Jerome Naylor, Nicole Stange-Thomann, Robert Barrett, Sante Gnerre, Michael Kamal, Manolis Kamvysselis, Evan Mauceli, Cord Bielke, Stephen Rudd, Dmitrij Frishman, Svetlana Krystofova, Carolyn Rasmussen, Robert L. Metzenberg, David D. Perkins, Scott Kroken, Carlo Cogoni, Giuseppe Macino, David Catcheside, Weixi Li, Robert J. Pratt, Stephen A. Osmani, Colin P. C. Desouza, Louise Glass, Marc J. Orbach, J. Andrew Berglund, Rodger Voelker, Oded Yarden, Michael Plamann, Stephan Seiler, Jay Dunlap, Alan Radford, Rodolfo Aramayo, Donald O. Natvig, Lisa A. Alex, Gertrud Mannhaupt, Daniel J. Ebbole, Michael Freitag, Ian Paulsen, Matthew S. Sachs, Eric S. Lander, Chad Nusbaum & Bruce Birren. The genome sequence of the filamentous fungus Neurospora crassa, *Nature*, 2003, vol. 422, pp. 859–868.