YEAST VOL. **14:** 701–710 (1998)

Pombe: A Gene-finding and Exon-intron Structure Prediction System for Fission Yeast

TING CHEN[1]* AND MICHAEL Q. ZHANG[2]

[1]*Department of Computer Science, The State University of New York, Stony Brook, NY 11794-4400, U.S.A.*
[2]*Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724, U.S.A.*

Received 16 June 1997; revised 27 September 1997

A special program developed by the authors, called Pombe, identifies protein coding regions in the *Schizosaccharomyces pombe* genome. Linear discriminant analysis was applied to predict 5′-terminal, internal, 3′-terminal exons (coding-exon) and introns. The accuracy of the prediction was tested by cross verifications. The sensitivity, specificity and correlation coefficient for the internal exon prediction were 98·5%, 99·9% and 98·3% respectively at the nucleotide level. Open reading frames were studied and used to predict intron-less genes: 99·0% of such genes were identified with correct stopping sites. The gene structure was determined by dynamic programming and the prediction achieved 97·0% correlation coefficient at the nucleotide level. The program is available at http://clio.cshl.org/genefinder. © 1998 John Wiley & Sons, Ltd.

Yeast 14: 701–710, 1998.

KEY WORDS — gene recognition; linear discriminant analysis; dynamic programming

INTRODUCTION

As the genomic DNA sequencing is scaled up to the mega-base level, it becomes more and more important to locate genes by fast and reliable means. Several complex systems for predicting gene structure have been developed in the last few years. Current gene-finding systems, such as SORFIND (Hutchinson and Hayden, 1992), GeneID (Guigo *et al.*, 1992), GRAIL (Uberbacher and Mural, 1991; Xu *et al.*, 1994; Lopez *et al.*, 1994), GeneParser (Snyder and Stormo, 1993), FEX (Solovyev *et al.*, 1994), and MZEF (Zhang and Marr, 1994; Zhang, 1997), are mainly based on a statistical or neural network approach. Many of them have also been integrated with tools to search databases for similarities.

Efforts to predict genes and gene structure (Fickett and Tung, 1992; Fickett, 1995; Gelfand, 1995) have been made for more than a decade. Although the number of existing algorithms is large, they generally consist of (1) searching for exon boundaries and (2) identifying potential coding regions. A few tried to predict entire exon–intron structure by heuristic exon assembly and had limited success.

Fission yeast is an important model organism for the study of biological processes at the cellular level, in particular the regulation of the eukaryotic cell cycle. It has become even more important in comparative genomics due to the availability of the budding yeast genome sequence. Although both genomes are about 15 million nucleotide base pairs in size, *Schizosaccharomyces pombe* has only three chromosomes, while *Saccharomyces cerevisiae* has 16. The individual genes of each species are also very different from one another. About one-third of fission yeast genes contain introns, which are very rare in budding yeast.

Most software available to the public was developed for gene-finding in vertebrates. Although FEX has been trained to a program called FEXY to predict splicing sites and exons in yeast DNA sequences, direct intron detection and gene structure prediction were not incorporated. Similar to some other lower eukaryotes, most introns of fission yeast genes are short, and splicing

*Correspondence to: T. Chen, Department of Computer Science, The State University of New York, Stony Brook, NY 11794-4400, U.S.A.

CCC 0749–503X/98/080701–10 $17.50

occurs mainly in the intron-definition mode as opposed to the exon-definition in vertebrates (review by Krizman and Berget, 1993). In fact, the first fission yeast gene prediction was done by an interactive program called INTRON.PLOT (Zhang and Marr, 1994). In Pombe, we use dynamic programming to combine exon and intron predictions into gene structure and improve overall accuracy.

DATA SETS

A high quality data set is required for building a good prediction system. Public databases contain various types of errors, such as sequencing errors, human editing errors and typing errors. Many statistical measurements are very sensitive to these errors so it is important to remove them from the data set. Redundancy also causes trouble. A database may contain many similar genes, and these genes will bias the prediction to them while ignoring other more important but rare ones. So identifying similar genes is another necessary step of selecting a high quality data set. In the following sections, we discuss how our learning data are extracted from GenBank and checked for errors and similarities.

Data extraction

Database entries containing *S. pombe* genomic DNA sequences were taken from GenBank release 95.0, and only those with split coding regions were extracted for training. The training set had 131 entries.

Annotation checking

Incorrect annotations happen quite frequently in GenBank. Generally, there is no good way to correct all of them without checking the original publications. However, we can limit possible errors by assuming each entry with a correct annotation should satisfy:

- The initiation site is ATG.
- The donor site is GT.
- The acceptor site is AG.
- The stopping site is either TAA or TAG or TGA.
- No stop codon interrupts the open reading frames.
- The length of coding regions is a multiple of three.

We compared each entry sequence that did not satisfy the criteria with the sequence in the original published paper. The following entries were checked and corrected accordingly: S64907, SPGCH1, SPRHP6, SPU52080, SPVATPA and YSPRPIIS3. Three entries were discarded because references were not available, leaving 128 entries in the data set.

Redundancy removal

Similar genes discovered by different research groups exist redundantly in GenBank. We define two genes to be similar if they share more than 90% identical nucleotide base pairs. FASTA, a fast sequence comparison program, was used to identify them. Five groups of entries were found to be similar. We selected one entry from each group based on whether it was genomic, had more nucleotide base pairs, or was published more recently.

SOME FISSION YEAST CHARACTERISTICS

Several distinctive characteristics of fission yeast were discovered during the process of building the prediction system. For instance, the arrangement of genes is compact, with a density of about 2000 bp per gene. There are very few overlapped genes. Most complex genes have a short 5′-terminal exon and a long 3′-terminal exon. The length of introns is between 40 and 700 nucleotide bp and each intron seems to contain a fairly conserved branch-site 10–30 bp upstream of its acceptor site.

METHODS

Discriminant analysis

We applied linear discriminant analysis to identify patterns between two alternative classes, *C1* (sites, introns or exons) and *C2* (pseudo sites, pseudo introns or pseudo exons). Linear discriminant analysis provides a linear function that separates two classes while minimizing misclassification. Assuming a p-feature variable $\overline{x} = \{x_1, x_2, \ldots, x_p\}$ is given, then the linear discriminant function

$$y = \sum_{i=1}^{p} \alpha_i x_i$$

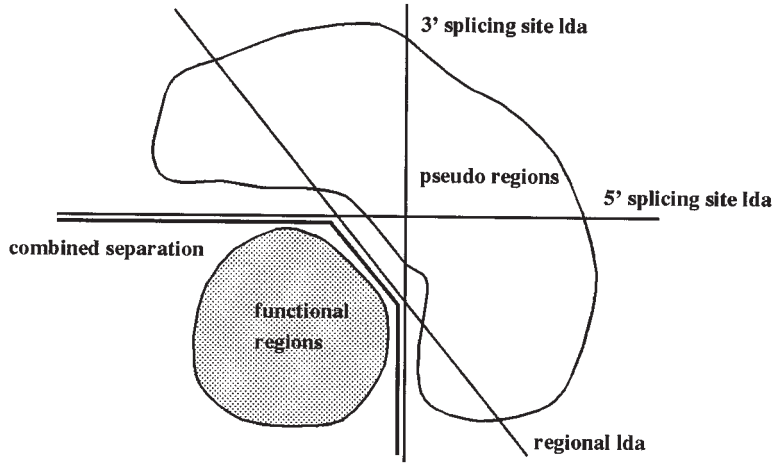


Figure 1. Approximating a non-linear two-class separation by three linear discriminant functions.

classifies $\bar{x}$ into class *C1* if $y \geq c$, and into class *C2* if $y < c$. The optimal selection of $\bar{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ and constant c can be determined by maximizing the ratio of between-class-variation to within-class-variation. If $\bar{\mu}_1$ is the sample mean vector for class *C1* and $\bar{\mu}_2$ is the sample mean vector for class *C2*, the optimal $\bar{\alpha}$ and c are given by

$$\bar{\alpha} = S^{-1}(\bar{\mu}_1 - \bar{\mu}_2)$$

and

$$c = \bar{\alpha}(\bar{\mu}_1 - \bar{\mu}_2)/2$$

where S is the pooled covariance matrix

$$S = \frac{S_1 + S_2}{n_1 + n_2 - 2}$$

and S_1 is the covariance matrix for class *C1* and S_2 is the covariance matrix for class *C2*, n_1 and n_2 are the number of samples for *C1* and *C2* respectively.

Exons and introns are recognized by two linear discriminations: initiation sites, donor sites and acceptor sites are identified before they are combined into introns and exons. Combination of linear functions is more powerful than a single linear function because it can approximate a non-linear discriminant function. Figure 1 shows a non-linear classification problem which can be approximated by three linear discriminant functions. Our exon and intron predictions are built on this model.

Oligonucleotide preferences

Oligonucleotide composition plays an important role in distinguishing sites and functional regions, for example, splicing sites, introns and coding regions.

Generally a pattern $p = p_1 p_2 p_3 \ldots p_l$ of length l is to be discriminated between two classes, C_0 and C_1. We can estimate the likelihood that pattern p belongs to class C_1 by the Bayesian method:

$$P(C_1|p) = \frac{P(p|C_1)P(C_1)}{P(p|C_1)P(C_1) + P(p|C_0)P(C_0)} = \frac{Pc_1(p)}{Pc_1(p) + Pc_0(p)}$$

where $P(C_1)$ and $P(C_0)$ are the *a priori* probabilities of two classes, C_0 and C_1. $P(p|C_1)$ and $P(p|C_0)$ are the *a posteriori* probabilities for pattern p to occur in class C_1 and class C_0; $P_{C_1}(p)$ and $P_{C_1}(p)$ are the frequencies of pattern p in class C_1 and class C_0. $P(C_1)$ and $P(C_0)$ are assumed to be equal: $P(C_1) = P(C_0)$. $P(C_1|p)$ is the preference of p in C_1.

To estimate the likelihood that a string $S = s_1 s_2 \ldots s_n$ belongs to class C_1 as against class C_0, we can average the preferences of all patterns $\{S_i\}$ of S:

$$\text{Score}(S) = \sum_{i=1}^{m} \frac{P(C_1|S_1)}{m}$$

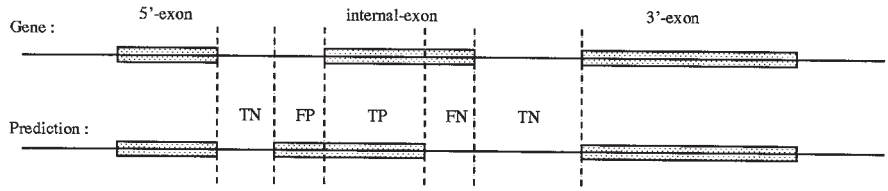


Figure 2. Measurement of internal exon prediction.

where S_i is the ith pattern of S and m is the number of patterns.

In our prediction system, the oligonucleotide preference of a DNA sequence S is measured by all of its b length substrings, called oligonucleotide compositions. Thus $S_i = s_i s_{i+1} \ldots s_{i+l-1}$ and $m = n - l + 1$. The oligonucleotide ($l = 6$) composition preferences are used to identify splicing sites, exons and introns. A variation of the formula is to pick a subset of $\{S_i\}$ as patterns for likelihood measurement, for example, in-frame hexamer preferences. In-frame hexamer preferences are used to predict coding regions and one in-frame hexamer nucleotide is exactly two adjacent codons.

Positional triplet preferences

Triplet composition of sequences adjacent to a particular site position may be used to discriminate such a site. We characterize functional boundaries, such as initiation sites, donor sites and acceptor sites, by the measurement of positional triplet preferences on a window around these sites.

Let $F^i_{t,k}$ and $F^i_{f,k}$ be the frequencies of a specific triplet k at the ith position of a window (L, R) in the true sites t and the pseudo sites f respectively. The triplet type k ranges from 1 to 64, representing all the possible triplets. Similar to the measurement in the oligonucleotide preferences, the probability of k at ith position belonging to a true site can be measured as

$$P_k(i) = \frac{F^i_{t,k}}{F^i_{t,k} + F^i_{f,k}}$$

For each splicing site, at some positions only certain triplets will appear. If the number of learning samples is not enough, we can just count those triplets who show significant difference in frequency between true sites and false sites. In a random sequence, each triplet is equally present at any position, thus $P_k(i)$ is 0·5.

The likelihood of a window (L, R) as a true site can be calculated by the following function:

$$\text{Score}(L,R) = \sum_{i=1}^{m} \frac{P_k(i)}{m}$$

where m is the number of triplets, and k is the triplet at ith position. Positional triplet preferences are used in the recognition of donor sites, acceptor sites and initiation sites.

Open reading frames

Open reading frames (ORFs) divide a DNA sequence into continuing, non-overlapped triplets, called codons. If an ORF appears in the coding region with the right frame, each codon will be translated into an amino acid. Three triplets, TAA, TAG and TGA, represent stop of translation. If a DNA sequence is random, on average, there is a stop codon for every 21 triplets (or 64 nucleotides). Therefore, the probability that an ORF has L triplets is $(1 - 1/21)^L \approx e^{-L/21}$. The probability of an ORF having 133 triplets (or 400 nucleotides) is less than 0·25%, which is highly significant.

For fission yeast, by assuming an ORF must contain a coding region if its length is above a certain threshold, we were able to identify many coding regions.

Measurement

In pattern recognition, typically distinguishing class 1 from class 0, the performance of a prediction system can be measured by the following statistics: true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

The internal exon prediction measurement on the nucleotide base pair level is shown in Figure 2, and the intron prediction measurement is shown in Figure 3. We did not consider regions outside the gene because these are either unknown or unreliable.

The accuracy of a prediction system is measured by sensitivity (Sn), specificity (Sp) and correlation co-efficiency (CC) as follows:

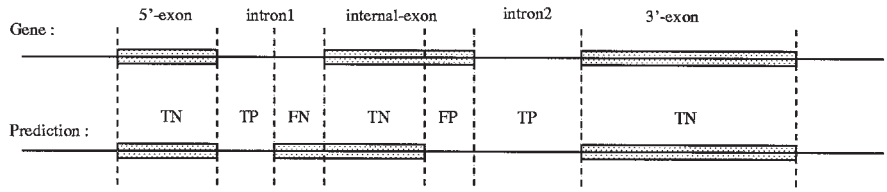
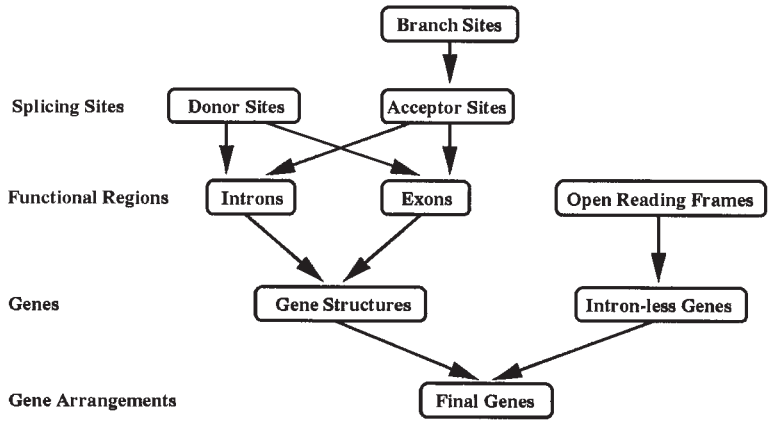


Figure 3. Measurement of intron prediction.



Figre 4. Prediction hierarchy.

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TP}{TP + FP}$$

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

RESULTS AND DISCUSSION

Prediction hierarchy

Figure 4 shows the hierarchy of the prediction. Splicing site linear discriminant functions were applied to predict sites before these sites were paired into regions and another level of LDA functions were used to identify functional regions. Finally these regions were assembled into a gene structure.

Splicing site prediction

A linear discriminant recognition system was built to distinguish true splicing sites from pseudo sites. Three features were measured: positional triplet preference, upstream hexamer preferences, and downstream hexamer preferences. We chose a threshold to retain all the true sites, and excess false positives would be dealt with through a second level linear discriminant system after these sites were assembled into introns or exons.

To identify donor sites, we extracted 261 true donor sites and 15,522 GT-containing pseudo sites from 116 genes in the data set. Three characteristics around these sites were used for developing and testing fission yeast donor site discriminant function. They were: the positional triplet preferences at the reserved region [− 5, +9]; the hexamer composition preferences in the potential coding region (upstream) [− 54, − 1]; the hexamer composition preferences in the potential intron region (downstream) [+1, +54]. All these features were observed to be statistically different in two classes.

Table 1. Optimal donor site linear discriminant function

	Total number	Classification: True sites	Classification: Pseudo sites
Donor sites	261	222	39
Pseudo sites	15,522	77	15,445

We built a linear discriminant function on these features. The optimal classification is shown in Table 1.

The LDA score of each GT site was calculated to range between 0 and 1. The minimum score for all 261 true donor sites was 0·09. In order to keep all the splicing sites, we chose the threshold 0·01, leaving 388 donor site candidates.

Similar calculations were done for acceptor sites and AG-containing pseudo sites. We extracted 261 acceptor sites and 16,140 pseudo sites. Four features were measured: the positional triplet preferences in the conserved region [− 18, +3], the hexamer composition preferences in the potential upstream intron region [− 54, − 1], the hexamer composition preferences in the potential downstream coding region [+1, +54], and the branch-site consensus preferences in the window [− 30, − 5] upstream of AG.

An important feature of a correct acceptor splicing site is the branch-site, which has a conserved consensus pattern and is roughly located in the region [− 30, − 5], upstream of the acceptor site. Using the weight matrix in Zhang and Marr (1994), we found the putative branch-site of each intron with the highest score, and aligned them to generate a new eight-nucleotide base pair matrix, shown in Table 2.

The branch-site preference was scored by the preference of a branch-site consensus vs a random 8-tuple. The actual position of the consensus is unknown and varies among different genes, so we scored all the 8-tuple nucleotide base pairs within the window [− 30, − 5] upstream of the AG site and selected the highest score as the putative branch-site preference.

We built a linear discriminant function to distinguish the acceptor sites from the pseudo sites. The optimal classification is shown in Table 3.

The LDA score of each AG site was calculated to range between 0 and 1. The minimum score for all 261 true acceptor sites was 0·002. We chose threshold 0·001 for a candidate acceptor site and we got 937 of them.

Table 2. Putative branch-site consensus

Position	A	C	G	T
1st	13·64	0·00	0·00	86·36
2nd	3·03	0·00	0·00	96·97
3rd	75·76	0·00	12·12	12·12
4th	0·00	100·00	0·00	0·00
5th	0·00	0·00	0·00	100·00
6th	96·97	0·00	3·03	0·00
7th	100·00	0·00	0·00	0·00
8th	0·00	78·79	0·00	21·21
Random	30·25	19·04	20·78	29·94

Table 3. Optimal acceptor site linear discriminant function

	Total number	Classification: Acceptor sites	Classification: Pseudo sites
Acceptor sites	261	248	13
Pseudo sites	16,140	50	16,090

Splicing site linear discriminant analysis identified more than 95% of the pseudo sites. The remaining 5%, along with all the true sites, were kept into the next level predictions. This largely eliminated a combinatorial explosion of intron or exon sample space due to large number of pseudo splicing sites.

Intron prediction

All the donor site candidates and the acceptor site candidates selected from the splicing site discriminant functions were paired as boundaries of potential introns. Unlike exons, which have a restriction of at least one ORF, introns do not have a general known constraint. However, analysis showed that all introns lie within the length range of [30, 700]. A possible explanation of the lower bound may be the physical hindrance constraints.

We paired 388 donor site candidates and 937 acceptor site candidates into regions within the above length constraint and obtained 1478 intron candidates. 261 of them were correct introns (with

Table 4. Optimal intron linear discrimination function

	Total number	Classification	
		Introns	Pseudo introns
Introns	261	177	84
Pseudo introns	1217	53	1164

Table 5. Intron and base pair level measurement of intron classification by excluding overlap

Measurement	Sn	Sp	CC
Intron level	0·929	0·963	
Base pair level	0·932	0·996	0·959

Sn, sensitivity; Sp, specificity; CC, correlation efficiency.

correct boundaries) and the rest either overlapped true introns or did not. As we have mentioned above, all the boundaries (sites) were selected based on their high scores, and some of the pseudo splicing sites may have higher scores than some of the true splicing sites. So, in addition to the splicing site features, the compositional hexamer preferences were added for intron discrimination. The compositional hexamer preferences of a region measure the difference between the compositional hexamer frequencies of introns and the compositional hexamer frequencies of exons.

The final intron prediction combined all the features of donor sites, acceptor sites, and the compositional hexamer preferences. The optimal discriminant function had the performance in Table 4.

Although we identified only 177/261 = 67·8% true introns (*sensitivity*) and 177/(177 + 53) = 77·0% correct classification (*specificity*), we observed that most of the pseudo introns, which were classified as putative introns, largely overlapped the true introns. An easy way to improve our results is to lower the threshold and exclude overlaps of introns. Only the highest score intron survived if two or more intron candidates overlapped. This idea is consistent with the intron-definition mode gene expression of yeast cell, because the cell is able to identify introns without ambiguity.

We tested our classification by cross validations on five test sets. Each set of tests was constructed in this way: among all training sequences, randomly select 80% for building a linear discriminant function and use it to predict introns on the remaining 20% of sequences.

The average *sensitivity*, *specificity*, and *correlation coefficiency* are shown in Table 5. On the intron level, 92·9% of introns were correctly predicted and 96·3% of predictions were introns. On the base pair level, 93·2% of intron base pairs were correctly predicted and 99·6% of predicted base pairs were intron base pairs. The base pair level *correlation coefficiency*, 0·959, is higher than other existing systems, to our knowledge. The linear discriminant function is very robust.

Internal exon prediction

As with intron discriminant analysis, internal exons also have flanking splicing boundaries: the acceptor splicing sites at the 5′-end and the donor splicing sites at the 3′-end. Potentially all of the selected donor site and acceptor site candidates can be paired to form exon boundaries. However, there are ORF constraints on the exon region, namely there must exist at least one ORF for each exon. Besides the splicing site and the ORF constraints, exons have a strong bias to in-frame hexamer frequencies. These in-frame hexamer frequencies are very important to identify long stretch exons but are much less sensitive to short exons, which are difficult to detect. So we added the length of exons as a separate feature in the discriminant analysis.

The number of internal exons in a gene is one less than the number of introns, and not every gene has an internal exon. In our learning data set, there are only 145 internal exons. If we paired 388 selected donor site candidates and 737 selected acceptor site candidates, we got another set of 374 pseudo exons, each with at least one ORF.

The in-frame hexamer preferences were measured between exon in-frame hexamer frequencies and intron compositional hexamer frequencies. The final internal exon prediction combined all the features of donor sites and acceptor sites, with the in-frame hexamer preferences and the log of exon length. The optimal discriminant function had the performance in Table 6.

If a true exon scores very high, the pseudo exons overlapping to it will also score high, and sometimes even higher than some true exons in other regions. To eliminate these clustering pseudo exons, we lowered the threshold of the optimal

Table 6. Optimal internal exon linear discriminant function

	Total number	Classification	
		Exons	Pseudo exons
Exons	145	108	37
Pseudo exons	374	24	350

Table 7. Exon and base pair level measurement of intron classification by excluding overlap

Measurement	Sn	Sp	CC
Exon level	0·942	0·960	
Base pair level	0·985	0·999	0·983

Sn, sensitivity; Sp, specificity; CC, correlation coefficiency.

discriminant function and excluded overlap of exons, as we did in intron prediction. Only the highest score exon survives if two or more exon candidates overlap.

We tested our classification by cross validation on five test sets. Each set of tests was constructed in this way: among all training sequences, randomly select 80% for building a linear discriminant function and use it to predict introns on the remaining 20% of sequences.

The average *sensitivity*, *specificity*, and *correlation coefficiency* are shown in Table 7. On the exon level, 94·2% of exons were correctly predicted and 96·0% of predictions were exons. On the base pair level, 98·5% of intron base pairs were correctly predicted and 99·9% of predicted base pairs were intron base pairs. The base pair level *correlation coefficiency*, 0·983, is higher than other existing systems, to our knowledge.

Almost every exon statistic is higher than that of the introns. We observed that only a small set of fission yeast genes have internal exons, and moreover, each internal exon is surrounded by two regulated introns with strong statistics. Thus the internal exon prediction is better.

Initial exon prediction

We built a linear discriminant recognition system to classify initiaton sites (ATG) from pseudo sites. Similar to what we did on the splicing sites, three features were measured: positional triplet preferences, upstream hexamer preferences, and downstream hexamer preferences. We extracted 116 true initiation sites and 4204 pseudo sites and the linear discriminant function had the optimal classification. All 116 initiation sites were identified correctly, and among 4204 pseudo sites, 4188 were identified and 16 were misclassified as true initiation sites.

Table 8. The relationship between the length of ORFs and the number of ORFs, exons and genes

Length (≥bp)	No. of ORFs	No. of exons	No. of genes
120	1308	285	116
180	490	216	112
240	247	174	108
300	176	146	100
360	142	130	96
420	118	112	92
480	105	101	87
540	92	90	83
600	88	86	80
680	80	79	74
720	71	71	67

Fortunately, we did not miss any true initiation sites in this classification, and we had only 16 incorrect initiation sites. Since our intron prediction is very accurate, the initial exon can be identified by the following strategy. For each initiation site candidate, we checked its ORF and compared it with the predicted introns. If they overlapped, the candidate was predicted as a true site, and its ORF was predicted as an initial exon. This strategy can eliminate some pseudo initiation sites, and others can be later judged in gene assembly.

Open reading frames

As discussed above, ORFs are among the most important features to identify the location of genes. They can work independently without needing extra information. We searched all the ORFs in the data set (Table 8).

Table 8 showed a strong correlation between the length of ORFs and the real exons: the longer an ORF, the more likely it is to be an exon. If we set a threshold of 420 bp for an ORF to be the putative location of a gene, solely based on this, we were able to determine 112 exons and 92 genes. The gap between 116 ORFs and 112 exons is small.

Table 9. Final exon open reading frame discriminant analysis

	Total number	Classification	
		Coding ORFs	Non-coding ORFs
Coding ORFs	112	98	14
Non-coding ORFs	334	0	334

Table 10. Gene assembly statistics on gene, exon, intron and base pair levels

Measurement	Sn	Sp	CC
Gene level	0·784	1·00	
Exon level	0·920	0·946	
Intron level	0·942	0·957	
Base pair level	0·987	0·999	0·972

Sn, sensitivity; Sp, specificity; CC, correlation efficiency.

However, the actual gap is even smaller because one exon may have more than one ORF and some other genes may not be discovered but may be sequenced in the data.

Final exon prediction

Among 116 genes, 112 have the longest ORFs in the final exons. We built a linear discriminant function to identify true ORFs and a final exon would be found if a predicted intron overlapped a predicted ORF.

We set a threshold of 210 bp for ORFs and obtained 334 non-coding ORFs and 112 coding ORFs. The linear discriminant function had two features, the in-frame hexamer preferences and the log of ORF length. The optimal classification is shown in Table 9.

We observed that all the coding ORFs we classified were true but 14 coding ORFs were missed and most of them were shorter than 210 bp. Thus, we chose to change the length threshold to include more ORFs and let the gene assembly program decide which are true.

Gene assembly

Gene assembly is an option in our program. The assembly combined all the predictions of 5′, internal, 3′ exons and introns. Since we had the intron prediction, the gene assembly can be easily done through dynamic programming with four rules in the following order: (1) a gene must have at least one ORF; (2) a gene with more components of exons or introns has an advantage over a gene with fewer components; (3) a gene with an initial exon has an advantage over a gene without an initial exon; (4) a gene with a final exon has an advantage over a gene without a final exon.

There is no danger of assembling multiple genes into one because the intron prediction limits the size of a gene in two ways. One is the length constraint on introns and another is that if two exons were adjacent in a gene there should exist an intron joining them.

The algorithm first translated framed exons and introns into vertices in a graph, with weights as the scores of the linear discriminant function. An adjacent exon and intron pair was translated into a directed edge in the graph. Each path in the graph was weighted according to the four rules described above. Dynamic programming was applied to find all the non-overlapped maximal weight paths, each of which gave the structure of a gene.

We tested the program in our 116 training data and obtained the statistics in Table 10.

We predicted 78% genes with correct splicing structure. The exon level and intron level prediction statistics are also very high and the base pair level correlation coefficiency is 0·972.

Intron-less genes

The intron-less genes were predicted by the same function for the final exons, except that we also considered any ORFs longer than 600 bp as genes. Also taken into account were ORFs that did not overlap with any introns and satisfied two rules: (1) the ORF has an ATG in the reading frame and (2) the distance from the first ATG to a stop codon is at least 360 bp.

So 204 intron-less genes were extracted from GenBank and they all had coding regions on the forward strand with correct boundaries and correct reading frames. We predicted 156 (76·5%) genes with correct initiation and stopping sites and 202 (99·0%) genes with correct stopping sites (and also reading frames). We had no false predictions. These 204 genes had not been used for training statistics.

Gene arrangements

Final output of gene arrangements combined gene assembly results with intron-less genes. The strategy was to fill the gaps between complex genes (with at least one intron) with intron-less genes. Overlapped genes on the same strand were not permitted, but we accepted genes with less than 20% overlaps on the different strands.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grant KO1 HG00010-05. We thank Barry Cohen and Steven S. Skiena, who reviewed the draft copy of this paper.

REFERENCES

Fickett, J. W. and Tung, C. S. (1992). Assessment of protein coding measures. *Nucl. Acids Res.* **20,** 6441–6450.

Fickett, J. W. (1995). ORFs and genes: how strong a connection. *J. Comput. Biol.* **2,** 117–123.

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

Guigo, R., Knuden, S., Drake, N. and Smith, T. (1992). Prediction of gene structure. *J. Mol. Biol.* **226,** 141–157.

Krizman, D. B. and Berget, S. M. (1993). Efficient selection of 3′-terminal exons from vertebrates DNA. *Nucl. Acids Res.* **21,** 5198–5202.

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

Lopez, R., Larsen, F. and Prydz, H. (1994). Evaluation of the exon prediction by GRAIL. *Genomics* **24,** 133–136.

Snyder, E. E. and Stormo, G. D. (1993). Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucl. Acids Res.* **21,** 607–613.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. (1994). Prediction of internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucl. Acids Res.* **22,** 5156–5163.

Uberbacher, E. and Mural, R. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Nat. Acad. Sci. USA* **88,** 11,261–11,265.

Xu, Y., Mural, R. J. and Uberbacher, E. C. (1994). Constructing gene models from accurately predicted exons: An application of dynamic programming. *Comput. Appl. Biosci.* **10,** 613–623.

Zhang, M. Q. and Marr, T. G. (1994). Fission yeast gene structure and recognition. *Nucl. Acids Res.* **22,** 1750–1759.

Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Nat. Acad. Sci. USA* **94,** 565–568.

