Plant Molecular Biology **37:** 803–806, 1998.

Identification of protein-coding regions in *Arabidopsis thaliana* genome based on quadratic discriminant analysis

M.Q. Zhang
Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, NY 11724, USA

Received 21 November 1996; accepted in revised form 6 February 1998

Key words: exon prediction, quadratic discriminant analysis, *Arabidopsis thaliana*

Abstract

A new method (MZEF) for predicting internal coding exons in genomic DNA sequences has been developed. This method is based on a prediction algorithm that uses the quadratic discriminant function for multivariate statistical pattern recognition. With improved feature measures, an *Arabidopsis thaliana*-specific implementation of MZEF is completed and made available to the plant genome community.

Marked by speedy identification and localization of complex disease genes [1, 2], biology has entered into a new era of genomics which has far reaching consequences in our understanding of life in nature [3]. As the Human Genome Project enters its large-scale sequencing phase, computational gene identification has become extremely important [4]. In an effort to improve the accuracy of exon prediction and to make a new tool freely available to the genome community locally in a timely fashion, a new program called MZEF (Michael Zhang's Exon Finder) was developed for identification of protein-coding regions in the human genome [5]. It is based on the quadratic discriminant analysis (QDA). Substantial improvements have been made when compared with existing methods: HEXON [6] (based on linear discriminant analysis) and GRAIL2 [7] (based on neutral networks). In a recent review [8], MZEF was ranked as the top algorithm for identification of human internal coding exons. In order to meet the need of the first plant genome sequencing project [9] and to facilitate world-wide gene-hunting effort, I have analyzed the statistical characteristics of *Arabidopsis thaliana* genome, redesigned the discriminant measures and implemented an *A. thaliana*-specific MZEF at the requests of many plant molecular biologists.

QDA (see e.g. [10]) is a powerful statistical multivariate pattern-recognition method. It may be thought of as a direct extension of the classical LDA (linear discriminant analysis) method pioneered by R.A. Fisher sixty years ago [11]. In general, a discriminant analysis can provide an optimal classification rule (in the sense of minimizing known errors) for discrimination of one population against another (in our case it would be for discrimination of real exons against pseudoexons). Graphically viewing the two populations as swarms of points in a multidimensional (feature) space, QDA can provide a more effective (curved) boundary between two swarms that have different co-variance structures than LDA which could only provide a straight (plane) boundary [12].

To assure the quality of data, 142 genomic sequences of the non-redundant data set, which had been carefully cleaned [13], was used. An internal coding exon candidate is defined as AG+ORF+GT (with 60 bp flanking sequence on each side), there were 110 848 samples taken from the region between the first coding exon and the last, which included 590 real exons and 110 258 pseudoexons. 10 feature variables were chosen for the discrimination. These 10 variables measure the following: exon length, upstream-intron score, branch-site score, 3′ss score, exon score, strand score, frame score, 5′ss score, downstream-intron score and GC ration (see Appendix for the definitions). The first 9 feature variables had been proved to be very effective in vertebrate exon predictions [1]. Although the branch sites in plant introns lack a strong consensus found in metazoan and the criteria for branch

Table 1.

$N = 22\,169$	I	II	III	IV	V	VI	VII	VIII	IX	X	Mean	SD
tp	107	105	101	95	107	97	103	104	105	100	102.4	4.1
fp	0	1	3	2	1	4	1	0	2	0	1.4	1.3
fn	11	13	17	23	11	21	15	14	13	18	15.6	4.1

site selection in plant was shown to be more relaxed in some genes [14], more rigorous statistical analysis did reveal a consensus WWCTRAW for *A. thaliana*, this signal, albeit weak, was still useful for improving acceptor site prediction (data not shown, see also [18]. This is consistent with the belief that basic splicing mechanism is conserved throughout eukaryotes. The addition of the last feacture variable, the GC ratio between ORF and flanking region, was motivated by the importance of AU-rich character of intron in dicots (see [14, 15] for references). The influence of AU-rich regions has been demonstrated by inserting AU-sequences at various places in a synthetic GC-rich intron thereby restoring its spliceability in dicot plants [16].

10 cross-validations were done as follows: I randomly selected 20% (from each population) as a test set and used the remaining to train QDA parameters (means and covariance matrix which determine the optimal classification surface). The result is shown as in the table (see Appendix for the notations).

This corresponds to, on average, $sn = 0.87$, $sp = 0.99$ and $cc = 0.92$. We see that QDA tends to have very high specificity. Although one could lower the threshold to increase the sensitivity at the expense of reducing the specificity, it is more desirable to have relatively high specificity (hence, less false positives) in practice, because it would allow bench scientists design less probes with higher confidence.

Most recently, a neural-network prediction system (NetPlantGene) has been developed and a study of splice site prediction in *A. thaliana* pre-mRNA was reported [17]. When compared with other programs, the overall performance of the coding/non-coding network ensemble of NetPlantGene on the test set was 0.76 in terms of the correlation coefficient (see the definition in Appendix; basically it is a single statistical measure achieving an optimal balance between sensitivity and specificity, it ranges from 0 for a random prediction to 1 for a 100% accurate prediction) as opposed to 0.55 for GeneMark [19] and it also outperformed Genefinder [20] and Grail [7] on splice site predictions. As NetPlantGene is not an exon prediction program, we suggest people should use both NetPlantGene and MZEF in parallel to achieve better results[1].

Currently, the genome database is a rapidly moving target. It goes without saying that any statistical rule-based method will depend on the training data set available at the time. The present data set may be biased, due to the way it has been generated (towards the genes which were the most abundant, the most expressed, the most easy to isolate or the most studied), a more representative sample will certainly be necessary in order to incorporate novel gene information. As more detailed understanding of splicing mechanism become available, better features variables will also be discovered. We plan to work closely with branch-scientists and with the genomic sequencing groups in order to further improve the accuracy.

MZEF is available at the anonymous ftp site phage.cshl.org in the directory pub/science/mzef (the author may be contacted at mzhang@cshl.org). It is also available through the World Wide Web at the URL of http://www.cshl.org/genefinder. The default parameters are set so that they optimized the total prediction at the base-pair level ($SN = 0.95$, $SP = 0.99$, $CC = 0.941$, which correspond to $sn = 0.88$ and $sp = 0.92$ at the exon level). One is referred to the README file and [1] for more technical descriptions. For convenience, the URLs for the other programs are also listed: http//www.cbs.dtu.dk/NetPlantGene.html for NetPlantGene, http: //compbio.ornl.gov/Grail-1.3/ for GRAIL, http://CCR-081.mit.edu/GENSCAN.html for

[1] As it took more than a year for this communication to be reviewed, there have been more plant gene-finding programs available. Most recently, Parnell *et al.* reported their analysis of the success of different gene prediction programs in identifying exons in the *A. thaliana* genome [21]. Here is a quote from their comparisons: 'The genomic copies of 25 cDNA identified by sequencing over 2 Mb of the *A. thaliana* genome served as standards to judge the ability of five different gene prediction programs to identify exons. MZEF, GRAIL, [7] and GenScan [22] identified over 70% of the possible exons, with MZEF scoring the highest success rate, FGENEA and FEXA [23] were less successful (Success rate: MZEF= 79.6%, GRAIL = 74.4%, GenScan = 72.2%, FGENEA = 54.9% and FEXA = 40.5%). In practice, using at least three programs is highly recommended.

GeneScan and http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html/for FGENEA and FEXA.

Acknowledgements

The author would like to thank Dr L. Parnell for sharing the comparison result before publication. This work is supported by a genome grant from the National Institutes of Health.

References

1. Wooster R *et al.*: Identification of the breast cancer susceptibility gene BRCA2. Nature 378: 789–792 (1995).
2. Tartaglia LA *et al.*: Identification and expression cloning of a leptin receptor, OB-R. Cell 83: 1263–1271 (1995).
3. Editorial: Capitalizing on the genome. Nature Genet 13: 1–5 (1995).
4. Collins F, Galas D: A new five-year plan for the U.S. Human Genome Project. Science 267: 43–46 (1993).
5. Zhang, MQ : Identification of protein coding regions in the human genome based on quadratic discriminant analysis. Proc Natl Acad Sci USA 94: 565–568 (1997).
6. Solovyev VV, Salamov AA, Lawrence CB: Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucl Acids Res 22: 5156–5163 (1994).
7. Uberbacher EC, Mural RJ: Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci USA 88: 1261–1265 (1991).
8. Claverie J-M: Computational methods for the identification of genes in vertebrate genomic sequence. Hum Mol Genet 6: 1735–1744 (1997).
9. Kramer D: First plant genome sequencing planned. Nature 383: 208 (1996).
10. McLachlan GJ: Discriminant Analysis and Statistical Pattern Recognition, John Wiley, New York (1992).
11. Fisher RA: The use of multiple measurements in taxonomic problems. Ann Eugen 7: 79–188 (1936).
12. Krzanowski WJ: Principles of Multivariate Analysis, p. 347. Clarendon Press, Oxford (1993).
13. Korning PG, Hebsgaard SM, Rouze P, Brunak S: Cleaning the GenBank *Arabidopsis thaliana* data set. Nucl Acids Res 24: 316–320.
14. Wiebauer K, Herrero J-J, Filipowicz W: Nuclear pre-mRNA processing in plants: distinct modes of 3′-splice-site selection in plants and animals. Mol Cell Biol 8: 2042–2051 (1988).
15. Waigmann E, Barta A: Processing of chimeric introns in dicot plants: evidence for a close cooperation between 5′ and 3′ splice sites. Nucl Acids Res 20: 75–81 (1992).
16. Goodall GJ, Filipowicz W: The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell 58: 473–483 (1989).
17. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S: Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucl Acids Res 24: 3439–3452 (1996).
18. Tolstrup N, Rouze P, Brunak S: A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites. Nucl Acids Res 25: 3159–3163 (1997).
19. Borodovsky M, Mclninch JD: GENMARK: Parallel gene prediction for both DNA strand. Comput Chem 17: 123–133 (1993).
20. Green P: Genefinder. Unpublished.
21. Parnell L, Dedhia N, McCombie WR: A statistical analysis of the success of exon prediction algorithims. The 1997 Biolotechnology Conference on the Arabidopsis Genome: From Sequence to Function, Cold Spring Harbor, NY, 11–14, Dec. 1997.
22. Burge C, Karlin S: Prediction of complete gene structure in human genomic DNA: J Mol Biol 268: 1–17 (1997).
23. Solovyev V, Salamov A: The Gene-Finder computer tools for analysis of human and model organisms genome sequences. ISMB 5: 294–302 (1997).

Appendix

The 10 feature variables were selected by experiments to achieve a reasonable prediction. They are defined as follows: If f_A is some frequency found in group *A*, we define the preference for *A* vs. *B* (say, exons vs. pseudoexons) as the ratio $f_A/(f_A + f_B)$ [6]. The first 9 feature variables are defined as follows (where $< \cdot >$ means an average, and splice site boundary is defined as (−1,1)): (1) exon-length, $x_1 = \log 10$ (actual length in bp); (2) exon-intron transition, $x_2 =<$ (intron hexamer frequency preference in the 54 bp window to the left of the 3′ss) $> - <$ (exon hexamer frequency preference in the 54 bp window to the right of the 3′ss) $>$; (3) branch-site score $x_3 =$ maximum branch score (measured by the putative log-likelihood score [18] in the window (−54,−3); (4) 3′ss score, $x_4 =$ position dependent triplet frequency preference for true 3′ss vs. pseudo-3′ss in the window (−24,3); (5) exon score, $x_5 =<$ (hexamer frequency preference for exon vs. intron) $>$; (6) strand score, $x_6 =<$ (hexamer frequency preference for the forward strand vs. the reverse strand) $>$: (7) frame score, $x_7 = max_{i=1,2,3}$ (frame specific hexamer frequency preference for exon vs. intron in frame *i*); (8) 5′ss score, $x_8 =$ positional dependent triplet frequency preference for true 5′ss vs. pseudo-5′ss in the window (−3,8); (9) intron-exon transition, $x_9 =<$ (exon hexamer frequency preference in the 54 bp window to the left of the 5′ss) $> - <$ (intron hexamer frequency preference in the 54 bp window to the right of the 5′ss $>$. The last is the GC ratio measure, $x_{10} =$ (GC contents in the ORF)/GC content in the flanking regions).

Table 2. The performance measures are the standard [1]:

	Predicted positives	Predicted negatives
Actual positives	true positives (TP)	false negatives (FN)
Actual negatives	false positives (FP)	true negatives (TN)
Sensitivity	$SN = \frac{TP}{TP+FN}$	
Specificity	$SP = \frac{TP}{TP+FP}$	
Correlation coefficient	$CC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}}$	

The exon-level measures are in lower case and the base-pair-level measures are in upper case. Namely, *tp* is the number of real exons in the predicted exons, *TP* is the number of nucleotides in the overlapping region between the real exons and the predicted exons; *fp* is the number of false exons in the predicted exons, *FP* is the number of nucleotides in the predicted exons that do not overlap with the real exons; *fn* is the number of the missed exons, *FN* is the number of nucleotides in the missed exons that do not overlap with any predicted exons. The statistic: sensitivity, specificity and correlation are widely used in many statistical validation test.