

Evaluation of Gene-Finding Algorithms by a Content-Balancing Accuracy Index

Chun-Ting Zhang^{1,*} and Ren Zhang²

¹Department of Physics, Tianjin University, Tianjin 300072, China

²Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital,
Tianjin 300060, China

*Corresponding author. Phone: +86-22-27402987; Fax: +86-22-23358329;
E-mail ctzhang@tju.edu.cn;

Abstract

A content-balancing accuracy index, called q_9 , to evaluate gene-finding algorithms has been proposed. Here the concept of content-balancing means that the evaluation by this index is independent of the coding and non-coding composition of the sequence being evaluated. Since the coding and non-coding compositions are severely unbalanced in eukaryotic genomes, the performance of gene-finding algorithms is either over- or under-evaluated by the widely used accuracy indices, e.g., the correlation coefficient, due to the lack of content-balancing ability. Using the new accuracy index q_9 , seven gene-finding algorithms, FGENES; GeneMark.hmm; Genie; Genescan; HMMgene; Morgan and MZEF, were compared and evaluated. It is shown that Genescan is still the best one, but with $q_9 = 89\%$, averaged over the prediction for 195 sequences. In addition to the content-balancing ability, q_9 has the merit of having definition in all possible cases. It is also shown that the traditional specificity s_p carries important information on the performance of the algorithm being evaluated. The set of sensitivity s_n , specificity s_p and the accuracy q_9 constitutes a complete kit to evaluate gene-finding algorithms at nucleotide level. In addition, a graphic method to compare and evaluate gene-finding algorithms has been proposed, too. Its major advantage is that the overall performance of algorithms can be grasped quickly in a perceivable form. Additionally, the new accuracy index q_9 may be applied to evaluate the performance of weather forecast, clinical diagnosis, psychological examination and protein secondary structure prediction etc.

Introduction

Computer-aided gene recognition is one of the most important bioinformatics problems. It is particularly necessary for genome annotations, when more and more completed genome sequences in public databases are available. The issues of gene-finding have been tackled by many groups since 1980s. An important review paper needs to be mentioned, in which various gene-finding algorithms proposed before 1992 were compared and evaluated by a unified benchmark (1). Further development directions were pointed out (1). Since then, great progress of computer-aided gene-finding studies has been made. The rapid progress of various sequencing projects stimulates great interests in looking for good gene-finding algorithms. On the other hand, the fast development of Internet and computer science in the past decade also highly facilitates bioinformatics research. Some review papers on these issues have been published recently (2-5). For a comprehensive list of papers regarding this topic, readers may visit an excellent web site which contains an update database of related literatures, maintained by W. Li, at <http://linkage.rockefeller.edu/wli/>.

An important problem is to compare and evaluate these algorithms. In addition to the work mentioned above (1), the comparison and evaluation of various algorithms using 570 sequences were performed (6). Recently, a detailed comparison and evaluation on seven newly developed algorithms based on 195 sequences were finished. The 195 sequences were entered into GenBank after the algorithms were developed and trained (7). The seven algorithms evaluated were: FGENES (8); GeneMark.hmm (9); Genie (10); Genescan (11); HMMgene (12); Morgan (13) and MZEF (14).

A question that naturally arises is how to compare and evaluate gene-finding algorithms? In addition to the sensitivity and specificity widely used, a single accuracy index is frequently used to evaluate various algorithms. Note that the gene recognition at nucleotide level will fall into one of the four categories, i.e., the true positive (TP), true negative (TN), false positive (FP) and false negative (FN). TP is the number of coding nucleotides that is recognized as coding, whereas TN is the number of non-coding nucleotides that is recognized as a non-coding. Similarly, FP (FN) is the number of non-coding (coding) nucleotides that is recognized as coding (non-coding). It must be kept in mind that information is always lost in the process of going from four numbers to one, called accuracy. The key issue is that the less information that is lost, the better the accuracy. Historically, nine accuracy indices were proposed. The first seven were listed in the monograph of Schultz and Schirmer (15), of which the seventh is also called correlation coefficient (CC). The eighth is called approximate correlation (AC) (6) and the ninth is called q_8 , proposed recently (16). Unfortunately, all of these nine accuracy indices do not possess the content-balancing ability. The content-balancing ability means that the evaluation by the accuracy index is independent of the coding or non-coding composition of the sequence being recognized. To illustrate this point more clearly, let us consider a concrete example. A DNA sequence is composed of 10% coding and 90% non-coding nucleotides, i.e., the composition of non-coding is much more than that of coding. Suppose that all nucleotides are recognized as non-coding. Accordingly, $q_3 = 0.9$ ($q_3 \in [0, 1]$), $q_8 = 0.78$ ($q_8 \in [-1, 1]$) and CC has no definition in this case, due to the appearance of 0/0 in the expression. The prediction that all nucleotides are non-coding provides no coding information. Lacking the content-balancing ability, the performance of the algorithm is obviously over-evaluated. The evaluation by these indices is misleading in some cases. The aim of this paper is to put forward a novel accuracy index, called q_9 , to

overcome the shortcomings of the above accuracy indices. The new accuracy index q_9 has definition in all possible cases and possesses the content-balancing ability.

Material and Method

The HMR195 database (7) was compiled and used to compare the seven algorithms mentioned above. The database contains 195 sequences which were entered into GenBank after the seven algorithms were developed. Therefore, the possibility that the sequences were used in the training set and again used in the test set is excluded. This database possesses the following characteristics: it is composed of DNA sequences of human, mouse and rat genomes; the mean length of sequence is 7096 bp; and the average number of exons per gene is 4.86, etc. For a detailed description of the database, refer to (1). The database is available at <http://www.cs.ubc.ca/~rogic/evaluation/dataset.html>. The HMR195 database was used in this paper.

As mentioned above, there are four possibilities in gene recognition for a given DNA base. They are the true positive; true negative; false negative and false positive recognition, respectively. The fractions of the above four cases are denoted by w , x , y and z , respectively (15). Sometimes the numbers of bases in the four cases are represented by TP, TN, FN and FP, respectively, too (6,17). There are simple relations between the two sets of notations

$$w = TP/N, \quad x = TN/N, \quad y = FN/N, \quad z = FP/N, \quad N = TP + TN + FN + FP. \quad [1]$$

Therefore, $0 < w, x, y, z < 1$ and $w + x + y + z = 1$. Based on these simple equations, a new accuracy index q_8 to evaluate gene-finding algorithms has been proposed recently (16)

$$q_8 = 1 - 2\sqrt{\frac{y^2 + z^2}{(w + y)^2 + (x + z)^2}}, \quad q_8 \in [-1, 1]. \quad [2]$$

The correlation coefficient (CC), also denoted by q_7 (15), is a widely used index in the literature

$$q_7 = CC = \frac{wx - yz}{\sqrt{(w + y)(w + z)(x + y)(x + z)}}, \quad q_7 \in [-1, 1]. \quad [3]$$

One of the drawbacks of q_7 is that it has no definition in many important cases, of which two need to be mentioned: (i) a sequence without coding region, i.e., $w + y = 0$; (ii) a sequence without non-coding region, i.e., $x + z = 0$. On the other hand, q_8 has definition in all cases.

For example,

$$q_8 = \begin{cases} \frac{x - z}{x + z}, & w + y = 0, \\ \frac{w - y}{w + y}, & x + z = 0. \end{cases} \quad [4]$$

Nevertheless, both eqs. [2] and [3] do not have the content-balancing ability. A content-balancing accuracy index means that the evaluation by this index is independent of the coding or non-coding composition of the sequence being recognized. As shown previously, a

sequence is composed of 10% coding and 90% non-coding region, and all are recognized as non-coding. This result corresponds to $w = 0$, $x = 0.9$, $y = 0.1$ and $z = 0$. Consequently, $q_8 = 0.78$, whereas q_7 has no definition. Such recognition is nothing useful, because it does not provide any coding information. The result $q_8 = 0.78$ indicates that the recognition is obviously over-evaluated. Similar over-evaluation cases occur for q_7 , too. Refer to (15) and the discussion below.

To overcome this drawback, a content-balancing accuracy index is proposed. First of all, introducing two content-balancing weight parameters

$$\alpha = \frac{1}{2} \frac{1}{w+y}, \quad \beta = \frac{1}{2} \frac{1}{x+z}, \quad [5]$$

we define the content-balancing transform

$$w' = \alpha \times w, \quad y' = \alpha \times y, \quad x' = \beta \times x, \quad z' = \beta \times z. \quad [6]$$

Note that $w' + y' = 0.5$ and $x' + z' = 0.5$. By using such a content-balancing transform, the set of (w, x, y, z) is transformed into a new set of (w', x', y', z') , where the coding and non-coding compositions, i.e., $w' + y'$ and $x' + z'$, are forced to be identical. The new content-balancing accuracy index q_9 is derived from q_8 ,

$$q_9 \equiv 1 - 2 \sqrt{\frac{y'^2 + z'^2}{(w' + y')^2 + (x' + z')^2}}. \quad [7]$$

Substituting eqs. [5] and [6] into eq. [7], and considering the two special cases of eq. [4], we define the content-balancing accuracy index q_9 as

$$q_9 \equiv \begin{cases} \frac{x-z}{x+z}, & \text{if } w+y=0, \\ \frac{w-y}{w+y}, & \text{if } x+z=0, \\ 1-\sqrt{2} \sqrt{\left(\frac{y}{w+y}\right)^2 + \left(\frac{z}{x+z}\right)^2}, & \text{if } w+y \neq 0 \text{ and } x+z \neq 0. \end{cases} \quad [8]$$

Therefore, q_9 has definition in all cases. Note that $q_9 \in [-1, 1]$, too. The new accuracy index q_9 defined in eq. [8] is the core of this paper. Before going into discussions in the following sections, let us reconsider the example mentioned above. A sequence is composed of 10% coding and 90% non-coding region, respectively, and all bases are recognized as non-coding. Substituting $w=0$, $x=0.9$, $y=0.1$ and $z=0$ into the third formula of eq. [8], we find $q_9 = 1 - \sqrt{2} = -0.414$, which is much more reasonable.

Results and discussions

Comparison between q_9 and q_7

Usually, a gene-finding algorithm is evaluated by the sensitivity s_n and specificity s_p , too.

Statistically, they are defined as

$$s_n = \frac{w}{w+y}, \quad s_p = \frac{x}{x+z}. \quad [9]$$

Note that $w+y$ and $x+z$ are the realistic coding and non-coding compositions, respectively. Therefore, the sensitivity is the fraction of coding nucleotides that are correctly recognized as coding; whereas specificity is the fraction of non-coding nucleotides that are correctly recognized as non-coding. Similarly, we can define another two parameters, the positive predictive ratio s_+ and negative predictive ratio s_-

$$s_+ = \frac{w}{w+z}, \quad s_- = \frac{x}{x+y}. \quad [10]$$

Note that $w+z$ and $x+y$ are the predicted coding and non-coding compositions, respectively. Therefore, s_+ is the ratio of the coding nucleotides over the predicted ones, whereas s_- is the ratio of non-coding nucleotides over the predicted ones. In some literature, e.g., (6), s_+ was ever called “specificity”. The word “specificity” we talk about here corresponds only to the second formula in eq. [9]. Note that the sensitivity s_n and specificity s_p are invariant subject to the content-balancing transform

$$s_n = \frac{w}{w+y} = \frac{w'}{w'+y'}, \quad s_p = \frac{x}{x+z} = \frac{x'}{x'+z'}. \quad [11]$$

This nice property indicates that both s_n and s_p are composition-independent. They are reliable parameters useful to evaluate algorithms. Rather than s_n and s_p , both s_+ and s_- do not have the property.

Baldi et al. pointed out that simply removing the terms of 0/0 from the expression of AC causes discontinuity, so the use of AC is not encouraged (18). Therefore, we will focus on the comparison between q_9 and q_7 . First of all, for an ideal recognition, i.e., $w+x=1$, $\Rightarrow q_9 = q_7 = 1$. For the worst recognition, i.e., $w+x=0$, $\Rightarrow q_9 = q_7 = -1$. For a random recognition, $q_9 = q_7 = 0$. In general cases, $q_9 \neq q_7$. Second, both q_9 and q_7 are symmetric with respect to s_n and s_p . Third, q_7 has no definition in the cases of $w+y=0$, $w+z=0$, $x+y=0$ and $x+z=0$, respectively, whereas q_9 has definition in all cases. Fourth, the most important difference between q_9 and q_7 is that q_9 is invariant subject to the content-balancing transform

$$q_9(w', x', y', z') = q_9(w, x, y, z); \quad [12]$$

whereas q_7 is not

$$q_7(w', x', y', z') \neq q_7(w, x, y, z). \quad [13]$$

Because the actual eukaryotic genomes are rich in non-coding sequences, the coding and non-coding composition of the sequence being recognized is frequently out of balance

severely. Lacking the content-balancing ability, the performance of algorithms is either over- or under-evaluated by the accuracy index q_7 (see the discussion below). In some cases, rather than q_7 and q_9 , it is convenient to use q'_7 and q'_9 defined by

$$q'_7 = \frac{1+q_7}{2}, \quad q'_9 = \frac{1+q_9}{2}, \quad q'_7, q'_9 \in [0, 1]. \quad [14]$$

Apply q_9 to compare and evaluate the seven gene-finding algorithms

The accuracy index q_9 is applied to compare and evaluate the seven well-known gene-finding algorithms. They are FGENES (8); GeneMark.hmm (9); Genie (10); Genescan (11); HMMgene (12); Morgan (13) and MZEF (14). For a given algorithm, the value of q_9 for each sequence in the HMR195 database is calculated, and then averaged over all sequences. The average q_9 for each of the seven algorithms is listed in Table I. For comparison, the average CC (or q_7), s_n, s_p, s_+ and s_- are also calculated and listed in Table I, too. Note that the above five quantities have no definition for some sequences. The sequences with no definition are simply left out in the average calculation.

Table I Various average evaluation parameters for the seven algorithms

Algorithm	No. ^a	q_9	q_7	AC	s_n	s_p	s_+	s_-	u^c	No. ^b
FGENES	195	0.76	0.83	0.84	0.86	0.98	0.88	0.95	0.13	195-5
GeneMark.hmm	195	0.78	0.83	0.84	0.87	0.97	0.89	0.95	0.12	195
Genie	195	0.79	0.88	0.89	0.91	0.98	0.90	0.98	0.08	195-15
Genescan	195	0.89	0.91	0.91	0.95	0.98	0.90	0.99	0.03	195-3
HMMgene	195	0.85	0.91	0.91	0.93	0.99	0.93	0.99	0.06	195-5
Morgan	127	0.63	0.69	0.70	0.75	0.95	0.74	0.95	0.24	127
MZEF	119	0.49	0.66	0.68	0.70	0.97	0.73	0.96	0.32	119-8

^aThe number of sequences selected for evaluating each algorithm in the HMR195 database. The calculation of q_9 is based on all of these sequences, because q_9 has definition for all possible cases.

^bThe numbers of sequences used to calculate CC, AC, s_n , s_p , s_+ , s_- and u are different, because these indices have no definition for some algorithms and for a few sequences. For example, some indices for Genie have no definition for 15 sequences out of 195.

^cThe unbalanced parameter defined by $u = (s_p - s_n)/((s_p + s_n)/2)$.

Based on the results listed in Table I, the following points need to be mentioned. First of all, the difference of \bar{q}_7 and \bar{q}_9 , i.e., $\Delta \equiv \bar{q}_7 - \bar{q}_9$, for each algorithm is calculated. The difference Δ ranges from 0.02 (Genescan) to 0.17 (MZEF). The values of Δ are always greater than zero, indicating that all the seven algorithms are on average over-evaluated by q_7 . The algorithm MZEF was the most severely over-evaluated, and it dropped by 17 percentage points (from 66% to 49%). The least severely over-evaluated program was Genescan, which dropped by about 2 percentage points only, from 91% to 89%. Of the seven algorithms, only Genescan and HMMgene have the q_9 accuracy greater than 80%. The q_9 accuracy indices for the remaining five algorithms are all less than 80%. The lowest value of q_9 is 49% (MZEF). Although the performance of all the seven algorithms is on average over-evaluated, for some individual cases, it is under-evaluated. The relations between q_9 and q_7 calculated based on the recognitions for 192 sequences (Genescan) and 111 sequences (MZEF) are shown in Fig. 1 (a) and (b), respectively. It can be clearly seen from these plots that some are over- and some are under-evaluated by q_7 , based on the benchmark of q_9 . A question needs to be answered: in the case that there is a difference between the values of q_7 and q_9 , which one is more reasonable? To answer this question, consider two typical cases corresponding to

the two points in Fig. 1 (a). Case 1: The DNA sequence being recognized has the GenBank ID U17081 with 9009 bp. Case 2: GenBank ID AF037207 with 5491 bp. The coding structures of both sequences were predicted by Genescan 1.0. The results are listed in Table II. As we can see that 100% of coding (s_n) and 96.9% of non-coding (s_p) nucleotides were correctly recognized in case 1. On the contrary, only 94.5% of coding (s_n) and 96.8% of non-coding (s_p) nucleotides were correctly recognized in case 2. Our intuitive feeling tells us that the prediction for case 1 is better than that for case 2. Nevertheless, $q_7 = 0.761$ and 0.792 , for case 1 and case 2, respectively, indicating that the performance of prediction in case 2 is better than that in case 1, based on the index q_7 . On the contrary, $q_9 = 0.956$ and 0.910 for case 1 and case 2, respectively, indicating that the new index q_9 reflects the real performance of the algorithm correctly. This example shows that q_9 reflects the performance of the algorithm being evaluated more objectively than q_7 does, because q_9 has the content-balancing ability, whereas q_7 does not.

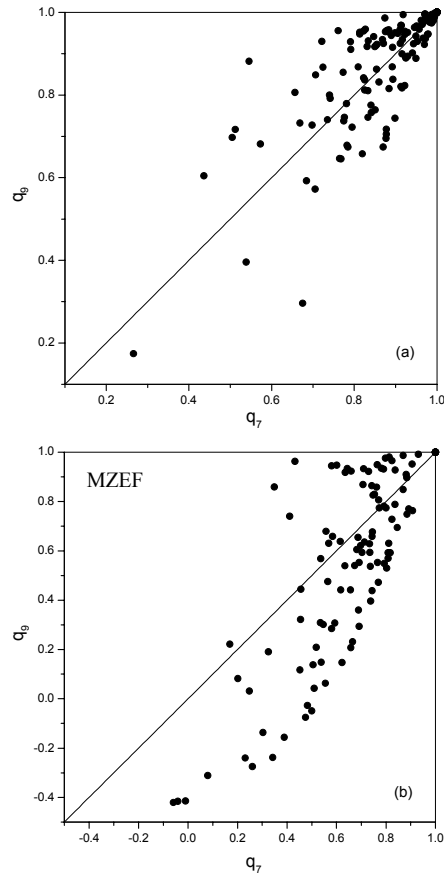


Fig. 1 Relations between q_9 and q_7 calculated based on the prediction for 192 sequences (Genescan) and 111 sequences (MZEF) are shown in Fig. 1 (a) and (b), respectively. It is clearly seen that some predictions are over- and some are under-evaluated by q_7 , based on the benchmark of q_9 .

Table II Gene prediction result for two sequences by Genescan 1.0

ID	w	x	y	Z	s_n	s_p	q_9	q_7	$w+y^a$
U17081	0.045	0.925	0.000	0.030	1.000	0.969	0.956	0.761	0.045
AF037207	0.066	0.900	0.004	0.030	0.910	0.968	0.910	0.792	0.070

^a $w + y$ is the content of coding nucleotides.

Second, because usually non-coding sequences are predominant in longer genomic regions, consequently, $x \gg z$, leading to the fact that s_p approaches to 1. Therefore, it was thought that the specificity defined in eq. [9] has high non-informative values. Based on this consideration, it was suggested to replace it by the positive predictive ratio s_+ (6). However, here we would like to show that s_p carries important information on the performance of the algorithm being evaluated. To illustrate this point, we define an unbalanced parameter $u = (s_p - s_n) / ((s_p + s_n) / 2)$. The average unbalanced parameter u for each algorithm is calculated and also listed in Table I. As we can see clearly that the unbalanced parameter u is negatively correlated with q_9 (with correlation coefficient equal to -0.99), indicating that the smaller the parameter u is, the better the accuracy q_9 is, or *vice versa*. Fig. 2 shows the relationship between q_9 and u . It is clearly seen that the seven points can be fitted well by a straight line, indicating that the correlation coefficient between q_9 and u is almost equal to -1. One feature of better algorithms, e.g., Genescan, is that it has smaller unbalanced parameter. On the contrary, the main feature of some relatively worse algorithms, e.g., MZEF, is that it possesses larger unbalanced parameter. Therefore, for some algorithms, e.g., MZEF, if the author re-adjusts the algorithm thresholds in the training process such that the unbalanced parameter is decreased, the accuracy would be raised remarkably. It is interesting to note that the extrapolation of the straight line in Fig. 2 to $u = 0$ results in a limit of accuracy, $q_{9\text{lim}} = 0.92$.

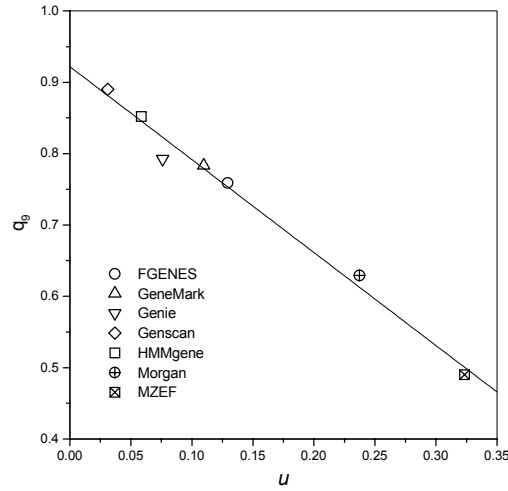


Fig. 2 Relation between the content-balancing accuracy index q_9 and the unbalanced parameter u for the seven algorithms. Note that the seven points associated with the seven algorithms, respectively, may be well fitted by a straight line. The correlation coefficient between q_9 and u is -0.99. The extrapolation of the straight line to $u = 0$ results in a q_9 value of 0.92.

In summary, the performance of all the seven algorithms is on average over-evaluated by the traditional correlation coefficient or q_7 . For individual cases, without having the content-balancing ability, the performance of gene-finding algorithms is either over- or under-evaluated by q_7 . Rather than q_7 , we believe that q_9 should be used to evaluate the performance of gene-finding algorithms in future studies. The specificity s_p carries important information on the performance of algorithms being evaluated. The set of s_n , s_p and q_9 constitutes an complete evaluation kit for an algorithm. One approach to raise the accuracy of current algorithms is to decrease the magnitude of the unbalanced parameter.

Graphic approach to evaluate gene-finding algorithms

Since $w' + x' + y' + z' = 1$, the four numbers can be mapped onto a point in a three-dimensional space (16). The coordinates of the mapping point are

$$\begin{cases} X' = 2(w' + x') - 1, \\ Y' = 2(w' + y') - 1, \\ Z' = 2(w' + z') - 1, \end{cases} \quad X', Y', Z' \in [-1, 1], \quad [18]$$

Because $w' + y' = 0.5, \Rightarrow Y' = 0$. Consequently, the four numbers, w', x', y' and z' , can be represented by a point in the two-dimensional $X' - Z'$ plane. Furthermore, the distribution region of mapping points is confined within a special region in the plane by the constraints $w' + y' = 0.5$ and $x' + z' = 0.5$. The constraint $x' < 0.5 \Rightarrow Z' \geq X' - 1$; $y' < 0.5 \Rightarrow Z' \geq -X' - 1$; $z' < 0.5 \Rightarrow Z' \leq X' + 1$ and $w' < 0.5 \Rightarrow Z' \leq -X' + 1$. Accordingly, the mapping points are confined within a regular square with its four vertices at (1,0), (0,1), (-1,0) and (0,-1) in the $X' - Z'$ plane, respectively. The ideal case, i.e., $w' + x' = 1$ corresponds to the point with coordinate (1,0). The worst case, i.e., $w' = x' = 0$ corresponds to the point with coordinate (-1,0). The closer a mapping point to (1,0) is, the better the recognition is. The closer a mapping point to (-1,0) is, the worse the recognition is. The x-axis divides the whole regular square into two triangles. In the upper triangle, $s_n > s_p$; whereas in the lower triangle, $s_n < s_p$. The y-axis divides the whole regular square into another two triangles. In the right triangle, $s_n + s_p > 1$, otherwise, in the left triangle, $s_n + s_p < 1$.

Therefore, based on the distribution of mapping points, the performance of algorithms may be evaluated by visual inspection. Fig. 3 (a) and (b) show the distributions of the mapping

points for the algorithms of Genescan and MZEF, respectively. Obviously, the former is better than the latter. The sensitivity and specificity of the Genescan algorithm are roughly balanced, indicating that the algorithm has higher accuracy. On the contrary, for the MZEF algorithm the distribution is unbalanced severely with respect to s_n and s_p (for most cases, $s_p > s_n$), indicating that the algorithm has lower accuracy. This observation is in good agreement with the calculation of the unbalance parameter u , shown in Table I and Fig. 2. Other algorithms can be evaluated and studied using the graphic method, too (data not shown here).

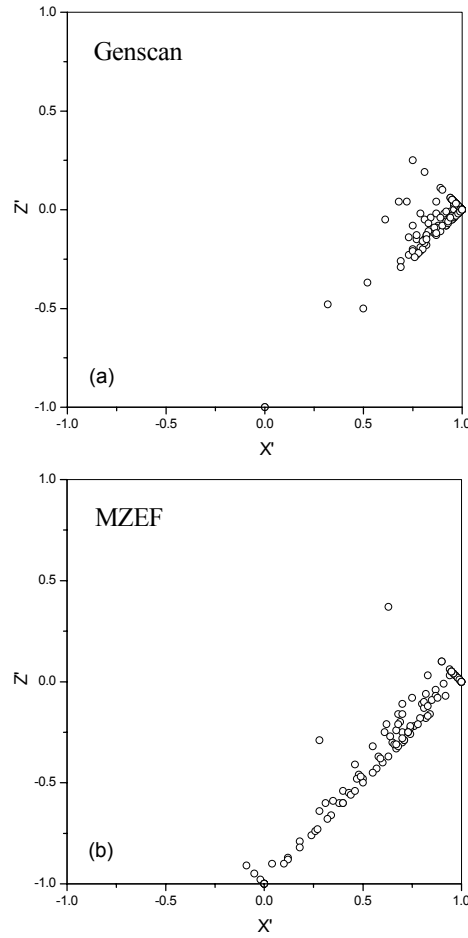


Fig. 3 Graphic analysis of gene prediction by two algorithms, (a) Genescan and (b) MZEF. The mapping points are confined within a regular square with its four vertices at (1,0), (0,1), (-1,0) and (0,-1) in the X' – Z' plane, respectively. The ideal case corresponds to the point with coordinate (1,0). The worst case corresponds to the point with coordinate (-1,0). The closer a mapping point to (1,0) is, the better the recognition is. The closer a mapping point to (-1,0) is, the worse the recognition is. The x-axis divides the whole regular square into two triangles. In the upper triangle, $s_n > s_p$; whereas in the lower triangle, $s_n < s_p$. Note that the distribution of the mapping points for Genescan is roughly symmetric with respect to $Z' = 0$ (Fig. 3 (a)), indicating that s_n and s_p are roughly balanced. Therefore, the algorithm has higher accuracy. On the contrary, the point distribution for MZEF in Fig. 3 (b) is severely unbalanced between s_n and s_p ($s_p \gg s_n$), indicating that the algorithm has relatively lower accuracy. One of the advantages of the graphic method is that the overall performance of the algorithm being studied can be grasped in a perceivable form.

Concluding Remarks

The correlation coefficient or q_7 is an accuracy index widely used in evaluating the performance of gene-finding algorithms at nucleotide level. In addition to the disadvantage that it has no definition in many important cases, q_7 lacks the content-balancing ability. Since the coding and non-coding compositions in most eukaryotic genomes are severely unbalanced, the performance of gene-finding algorithms is either over- or under-evaluated by the accuracy index q_7 . To solve this problem, the content-balancing accuracy index q_9 has been proposed here. Using the new index q_9 , seven recently developed gene-finding algorithms were compared and evaluated. The program of Genescan is still the best one, and the average q_9 is 89%. It is found that the sensitivity s_n and specificity s_p are severely unbalanced for most of the seven algorithms, leading to relatively lower q_9 accuracy. For most algorithms if their thresholds can be readjusted in the training process such that the unbalanced situation of s_n and s_p is improved, the q_9 accuracy would be remarkably increased. It is shown that the traditional specificity s_p carries important information on the performance of the algorithm being evaluated. The set of sensitivity s_n , specificity s_p and the accuracy q_9 constitutes a complete kit to evaluate gene-finding algorithms at nucleotide level. In addition, a graphic method to compare and evaluate gene-finding algorithms has been proposed, too. One of the advantages of this approach is that the performance of algorithms can be grasped quickly in a perceivable form. The new accuracy index q_9 has the merit of having definition in all cases. Besides q_7 , we believe that q_9 should be used to evaluate the performance of gene-finding algorithms at nucleotide level in future studies. In addition to

evaluate gene-finding algorithms, the new accuracy index q_9 could be used widely in all statistical areas where q_7 was or is being used. The applicable areas include evaluations of the performance of weather forecasts, clinical diagnoses, psychological examinations and protein secondary structure predictions etc.

Acknowledgements

We thank Dr. Rogic for kindly sending us the necessary data used in this work. The present study was supported in part by the 973 Project of China, grant no. G1999075606.

References and Footnotes

1. Fickett, J. W. and Tung, C.-S., *Nucleic Acids Res.* 20, 6441-6450 (1992).
2. Gelfand, M. S., *J. Comput. Biol.* 2, 87-115 (1995).
3. Fickett, J. W., *Comput. Chem.* 20, 103-118 (1996).
4. Claverie, J.-M., *Hum. Mol. Genet.* 6, 1735-1744 (1997).
5. Stormo, G. D., *Genome Res.* 10, 394-397 (2000).
6. Burset, M., and Guigo, R., *Genomics* 34, 353-367 (1996).
7. Rogic, S., Mackworth, A. K., and Ouellette, F. B. F., *Genome Res.* 11, 817-832 (2001).
8. Solovyev, V., Unpublished.
9. Lukashin, A. V., and Borodovsky, M., *Nucleic Acids Res.* 26, 1107-1115 (1998).
10. Kulp, D., Haussler, D., Reese, M. G., and Eeckman, F. H., In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology* (eds. D. States et al.), pp.134 –142. AAAI Press, Menlo Park, CA (1996).

11. Burge, C., PhD thesis. Stanford University, Stanford, CA (1997).
12. Krogh, A., In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (eds. T.Gaasterland et al.), pp.179 –186, AAAI Press, Menlo Park, CA (1997).
13. Salzberg, S., Delcher, A., Fasman, K.,and Henderson, J., J. Comp. Biol. 5, 667-680 (1998).
14. Zhang, M.Q., Proc. Natl. Acad. Sci. 94, 565-568 (1997).
15. Schultz, G. E. and Schirmer, R. H., Principles of Protein Structure. Springer, New York, (1979).
16. Zhang, C.-T. and Zhang, R., Proteins 43, 520-522 (2001).
17. Burge,C. and Karlin, S., J. Mol. Biol. 268, 78-94 (1997).
18. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. and Nielsen, H., Bioinformatics 16, 412-424 (2000).