BIOINFORMATICS

Vol. 16 no. 11 2000
Pages 960–967

Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences

Tetsuo Nishikawa [1,2,*], Toshio Ota [1,3] and Takao Isogai [1]

[1]Helix Research Institute, Chiba, Japan

Received on April 29, 2000; revised and accepted on May 31, 2000

Abstract

Motivation: *In the previous works, we developed ATGpr, a computer program for predicting the fullness of a cDNA, i.e. whether it contains an initiation codon or not. Statistical information of short nucleotide fragments was fully exploited in the prediction algorithm. However, sequence similarities to known proteins, which are becoming increasingly available due to recent rapid growth of protein database, were not used in the prediction. In this work, we present a new prediction algorithm based on both statistical and similarity information, which provides better performance in sensitivity and specificity.*

Results: *We evaluated the accuracy of ATGpr for predicting fullness of cDNA sequences from human clustered ESTs of UniGene, and we obtained specificity, sensitivity, and correlation coefficient of this prediction. Specificity and sensitivity crossed at 46% over the ATGpr score threshold of 0.33 and the maximum correlation coefficient of 0.34 was obtained at this threshold. Without ATGpr we found it effective to use alignments with known proteins for predicting the fullness of cDNA sequences. That is, specificity increased monotonously as similarity (identity of the alignments) increased. Specificity was achieved greater than 80% if identity was greater than 40%. For more effective prediction of fullness of cDNA sequences we combined the similarity (identity of query sequence) with known proteins and ATGpr score. As a result, specificity became greater than 80% if identity was greater than 20%.*

Availability: *The prediction program, called ATGpr_sim, is available at http://www.hri.co.jp/atgpr/ATGpr_sim.html*

Contact: *nisikawa@crl.hitachi.co.jp*

*To whom correspondence should be addressed.

[2]Present address: Biosystems Research Department, Central Research Laboratory, Hitachi Ltd, 1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo, 185-8601, Japan.

[3]Present address: Kyowa Hakko Kogyo Co. Ltd, Tokyo Research Laboratories, 3-6-6 Asahi-machi, Machida-shi, Tokyo, 194-8533, Japan.

Introduction

More than one million human cDNA fragment sequences have already been published by ESTs projects (Hillier *et al.*, 1996). The ESTs sequences are often incomplete in the 5′-region of full-length cDNA sequences. However, for functional analysis of gene, it is important to obtain clones including intact protein coding sequences (complete clones). To select them efficiently from given cDNA fragments, computer programs as well as effective methods for generating complete clones (Maruyama and Sugano, 1994) are required. We therefore previously developed a computer program, ATGpr (Salamov *et al.*, 1998), which estimated reliability of prediction by using statistical information; By ATGpr each ATG in a given DNA sequence was predicted as a true translation initiation codon or not. By using complete cDNA sequences, we evaluated the accuracy of the initiation codon prediction by ATGpr in the previous study. However, prediction of initiation codon is difficult in principle because of many false ATG codons contained in a cDNA sequence. Since ATGpr uses only statistical information derived from the cDNA sequences, its accuracy can be improved when information of other known proteins is added.

The purpose of this paper is to improve the ATGpr and to clarify how similarity such as identity contribute to fullness prediction of cDNA sequences. We have developed a new prediction method, which uses both statistical information and similarities with other known proteins to obtain higher accuracy of fullness prediction for fragment sequences of cDNA clones. Actually, lots of newly determined protein sequences from genome sequencing projects (microbial genome projects are listed at TIGR homepage, http://www.tigr.org/) are available. We used human UniGene data (Schuler, 1997) as a source of cDNA fragment sequences. First, we examined the prediction accuracy of ATGpr for fragment sequences in UniGene. Second, we evaluated (without ATGpr) the accuracy of the prediction that uses alignment information

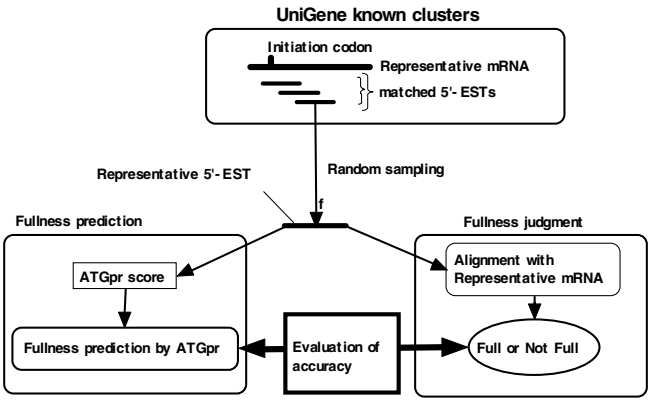

Fig. 1. Evaluating the accuracy of fullness prediction of cDNA sequences using UniGene. Prediction is by ATGpr.

of known proteins. Third, we investigated how to combine the ATGpr score with the alignment information for efficient prediction of fragment sequences, and we show high accuracy of our proposed prediction method using UniGene.

Method

Prediction of fullness of cDNA by ATGpr using UniGene

As shown in Figure 1, 5732 clusters (Full UniGene clusters), which were made by removing incomplete clusters without mRNAs with translation initiation codons from known human UniGene clusters (Build49, 6963 clusters; 'known' means that they include mRNA sequences), were used. A representative mRNA (the longest mRNA in a cluster) and 5′-ESTs were extracted from each full UniGene cluster. The 5′-ESTs were compared with the representative mRNA by using BLASTN (BLAST2.0), and the 5′-ESTs satisfying matching conditions (alignment length $\geqslant$200 bases; identity $\geqslant$90%, where identity was defined as the proportion of concordance at the same position in an alignment) were selected (4421 clusters). A 5′-EST was randomly sampled from each cluster (representative 5′-EST). Whether representative 5′-EST included an initiation codon was judged by using the alignment of the representative mRNA sequences. The number of representative 5′-EST which included an initiation codon was 840, and that of representative 5′-EST which did not include an initiation codon was 3581. The maximum ATGpr score in all ATGs included in each representative 5′-EST (we call this value the ATGpr score) was then calculated. When the ATGpr score was greater than a given threshold, the cDNA sequence was predicted as 'full'; that is, it includes an initiation codon.

Prediction of fullness of cDNA by using ATGpr score and similarity with other proteins

It has been empirically observed that N-terminals of similar sequence—proteins appear close to each other in their alignments. Figure 2 compares ANFB_HUMAN BRAIN NATRIURETIC PEPTIDE PRECURSOR (BNP) with OWL protein sequences, and several alignments are obtained. There are three homologous protein sequences to the query in the hit list. Their initiation codons are aligned at the same position. We therefore found that fullness can be predicted by aligning unknown cDNA fragments with known proteins (fullness means that they include an initiation codon). We call this prediction method 'prediction by similarity', and we developed the following procedure. An unknown cDNA sequence is aligned with known proteins by BLASTX as shown in Figure 3. When the alignment satisfies certain conditions and the not-aligned length of the 5′-terminal of the unknown cDNA is longer than that of the aligned protein multiplied by three, the unknown cDNA sequence is predicted as 'full'; that is, it includes an initiation codon. The alignment conditions are identity, consensus length and *E*-value. We evaluated this method from the following points:

- What is the optimal alignment condition?
- What is the accuracy of the prediction under the optimal alignment condition?
- What is the optimal combination of similarity and ATGpr on prediction?

A representative mRNA and hit 5′-ESTs are extracted from each full UniGene cluster (Figure 1) as shown in Figure 4. Whether a representative 5′-EST randomly sampled from each cluster includes an initiation codon is checked by using the alignment with the representative mRNA sequence. The representative 5′-EST is compared with protein sequences in the full-OWL protein database (152 308 entries); that is a protein database made by removing fragment sequences and sequences without methionin at N-terminal from the OWL protein database.

To evaluate the accuracy of the prediction by similarity, a set of alignment parameters, in this case the set of all possible identities of which definition is stated on the 3rd page, is divided into several disjoint subsets. For the prediction by similarity and ATGpr, the prediction space, which consists of identity and ATGpr score, is divided into many disjoint subspaces. Then fullness prediction in each subset or subspace is performed and the accuracy is evaluated, respectively.

```
Query    = Homo sapiens cDNA, 5' end ,  GenBank Accession=AA216138 ,  (371 bases)
           ANFB_HUMAN BRAIN NATRIURETIC PEPTIDE PRECURSOR (BNP)
Database= Full-OWL protein database (152,308 entries)

 Sequences producing significant alignments:                          Score  E      Id     Consensus length
1) pir|P16860|ANFB_HUMAN BRAIN NATRIURETIC PEPTIDE PRECURSOR (BNP)...    155  6e-3893%    87b          same as the query
2) pir|P16859|ANFB_CANFA BRAIN NATRIURETIC PEPTIDE PRECURSOR (BNP)...     73  5e-1355%    86b          homologous to the query
3) pir|P07634|ANFB_PIG BRAIN NATRIURETIC PEPTIDE PRECURSOR (BNP). ...     68  2e-1155%    77b          homologous to the query
4) pir|P13205|ANFB_RAT BRAINNATRIURETIC PEPTIDE PRECURSOR (BNP) (...     45  1e-04  42%    63b          homologous to the query
5) pir|U87996|CAU87996 CAU87996 NID: g2351701 - Candida albicans.        43  6e-0438%    88b          no relation to the query


 1)  Score=155 bits (388), Expect=6e-38, Identities=81/87 (93%), Positives=81/87 (93%), Gaps=3/87 (3%)
  Query: 100 MDPQTAPSRALLLLLFLHLAFLGGRSHPLGSPGSASDLETSGLQEQRNHLQGKLSELQVE 279
             MDPQTAPSRALLLLLFLHLAFLGGRSHPLGSPGSASDLETSGLQEQRNHLQGKLSELQVE
  Sbjct: 1   MDPQTAPSRALLLLLFLHLAFLGGRSHPLGSPGSASDLETSGLQEQRNHLQGKLSELQVE 60


 2)  Score=73.0 bits (176), Expect=5e-13, Identities=48/86 (55%), Positives=55/86 (63%), Gaps=7/86 (8%)
  Query: 100 MDPQTAPSRALLLLLFLHLAFLGGRSHPLGSPGSASDL----ETSGL---QEQRNHLQGK 258
             M+P  A  RALLLLLFLHL+ LGGR HPLG    AS+     E SGL   QE    L+
  Sbjct: 1   MEPCAALPRALLLLLFLHLSPLGGRPHPLGGRSPASEASEASGLWAVQELLGRLKDA 60


 3)  Score=44.9 bits (104), Expect=1e-04, Identities=27/63 (42%), Positives=37/63 (57%)
  Query: 100 MDPQTAPSRALLLLLFLHLAFLGGRSHPLGSPGSASDLETSGLQEQRNHLQGKLSELQVE 279
             MD Q    + +LLLLFL+L+ LGG SHPLGSP  +   E S +Q+    ++ K  E+
  Sbjct: 1   MDLQKVLPQMILLLLFLNLSPLGGHSHPLGSPSQSP--EQSTMQKLLELIREKSEEMAQR 58


 4) Score=42.6 bits (98), Expect=6e-04, Identities=34/88 (38%), Positives=43/88 (48%), Gaps=14/88 (15%)
  Query: 9   EGRWEANPDASQQQQQQ---------QKQQQQPPQSLQRHGSPDSTFPGAPAPALLASGF 161
             E + +   D+SQQQQQQ         Q+QQQQP Q L  H    S  P    P L +SG
  Sbjct: 442 EQQRQQRTDSSQQQQQQQKQHQYQQKSQQQQQQPQQPLSSHQGGTSHIPKQVPPTLPSSGP 501
```

Fig. 2. Coincidences of initiation codons between proteins at various similarity levels (Examples of BLASTX alignments).

Results

Prediction of fullness of cDNA by ATGpr

Specificity, sensitivity, and correlation coefficient of the prediction by using ATGpr are plotted over the threshold of ATGpr score from 0 to 1 as shown in Figure 5. The specificity, the sensitivity, and the correlation coefficient are defined as follows.

$$\text{Specificity} = \frac{a}{a+c}. \tag{1}$$

$$\text{Sensitivity} = \frac{a}{a+b}. \tag{2}$$

$$\text{Correlation coefficient} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{3}$$

where a is number of full sequences predicted as full, b is number of full sequences predicted as not full, c is number of not full sequences predicted as full and d is number of not full sequences predicted as not full.

Figure 5 shows specificity increases from 19% (threshold: 0) to 100% (threshold: 1) and sensitivity decreases from 100% (threshold: 0) to 0 (threshold: 1). Specificity and sensitivity cross at 46% over the threshold of 0.33. The maximum correlation coefficient is achieved approximately over the same ATGpr score as the crossing of specificity and sensitivity occurs. The maximum correlation coefficient is 0.34, and this is not so high. It is caused by that the fullness-proportion of UniGene is only 19%. Though specificity and sensitivity of 46% is not so high, it is much higher than 19% (at threshold: 0), which is the proportion of true full fragments in UniGene. At the threshold of 0.8, specificity is 80%, but sensitivity becomes 5%, which means that the proportion of true full cDNA fragments not predicted as full is much (95%). To increase sensitivity of the prediction keeping high specificity, using similarity of cDNA sequences with other proteins is deemed effective.

Prediction of fullness of cDNA by using similarity with other proteins

First, we evaluated the accuracy of the fullness prediction by using only similarity without ATGpr. Totally 4421 representative 5′-ESTs of known UniGene clusters are compared with the full-OWL protein database, and for each hit of the comparison the fullness prediction by similarity using the hit alignment was performed. Totally 11 950 hits were used for the prediction. In Table 1 we showed observed numbers of hits predicted full, predicted not full, originally full, and originally not full in each identity range under the conditions where the E-value $\leqslant 1$. To interpret these numbers we discuss the accuracy of the prediction by specificity and sensitivity as follows. In Figures 6a, b, and c, the range of identity of the alignment between an EST and a protein is divided into ten

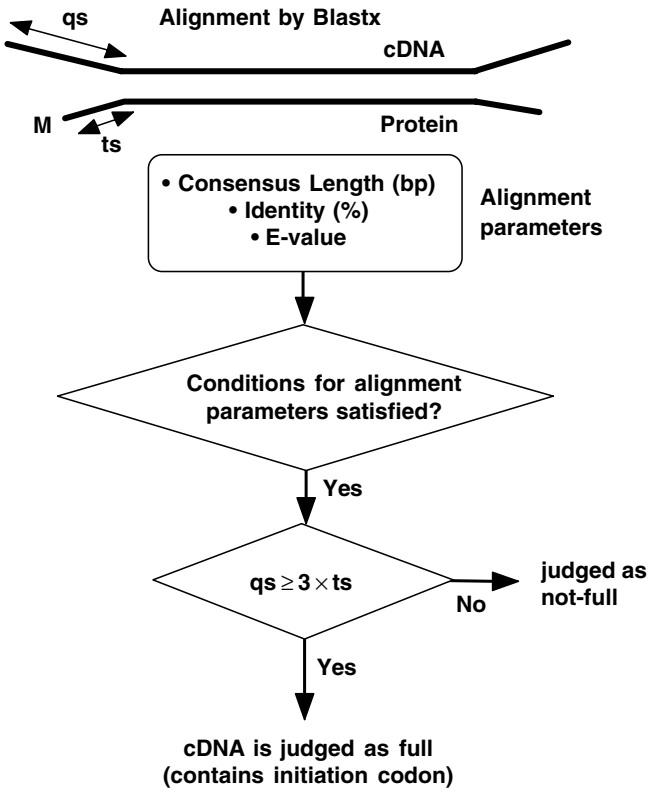

Fig. 3. Flow chart of fullness prediction by similarity with other proteins.

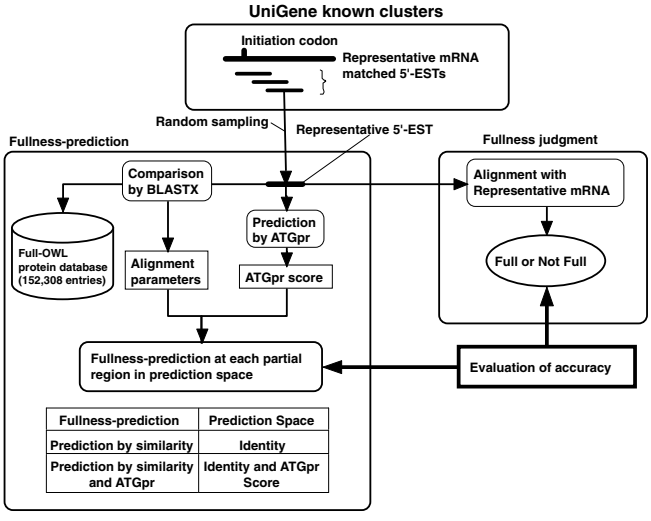

Fig. 4. Evaluating the accuracy of fullness prediction by a combination of similarity with other proteins and ATGpr using UniGene.

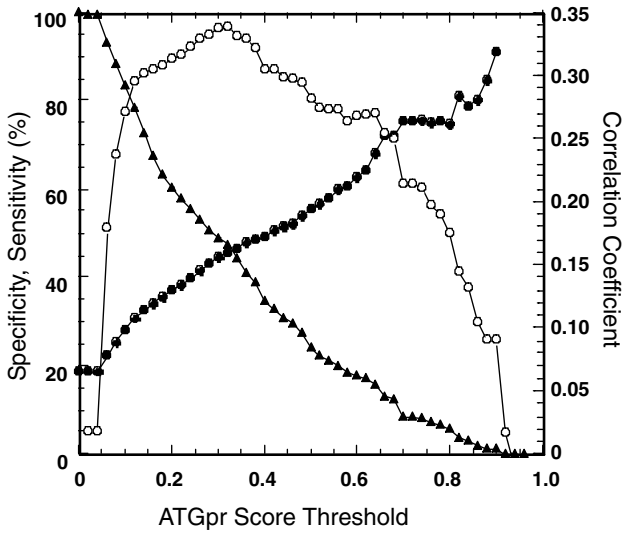

Fig. 5. Specificity and sensitivity and point correlation coefficient of the prediction of fullness of cDNA by ATGpr as a function of ATGpr score threshold. Specificity (●), Sensitivity (▲), Correlation coefficient (○).

disjoint subsets. Each figure contains lines representing E-values of $\leqslant 1$, $\leqslant 0.1$, $\leqslant 0.01$, and $\leqslant 0.000\,01$. The accuracy of specificity or sensitivity of the predictions by similarity in each identity range is plotted. Specificity of positive prediction, in other words, the proportion of sequences truly predicted as full to those predicted as full, is shown in Figure 6a. Sensitivity of positive prediction, in other words, the proportion of sequences truly predicted as full to entire true full sequences, is shown in Figure 6b. Sensitivity of negative prediction, in other words, the proportion of sequences truly predicted as not full to entire true not-full sequences, is shown in Figure 6c. In Figure 7, correlation coefficient of the prediction is shown. Figure 6a shows that specificity of positive prediction is an increasing function of identity and is greater than 80% over the identity greater than 40%. Specificity varies depending on E threshold over the identity of 20–40%. Figure 6b shows that sensitivity of positive prediction is an increasing function of identity and is greater than 80% over the identity greater than 50%. And Figure 6c shows that sensitivity of negative prediction does not depend on E threshold and is greater than 95%. Correlation coefficient is an increasing function of identity, and is greater than 0.85 over the identity greater than 50%. Figures 6a and b show that specificity and sensitivity of positive prediction are higher at lower E threshold over the identity range of 20–40%. But if we focus on the number of sequences, not on the proportion, we obtain opposite trend in terms of E-value.

Table 1. Observed numbers of hits predicted full, predicted not full, originally full, and originally not full in each identity region under the conditions where the E-value $\leqslant 1$

Identity (%)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100
Number of hits predicted full	0	0	43	247	304	312	291	366	353	541
Number of hits predicted not full	0	2	481	1867	1125	853	920	1000	1201	2044
Number of hits originally full	0	0	107	362	373	306	260	333	337	529
Number of hits originally not full	0	2	417	1752	1056	859	951	1033	1217	2056

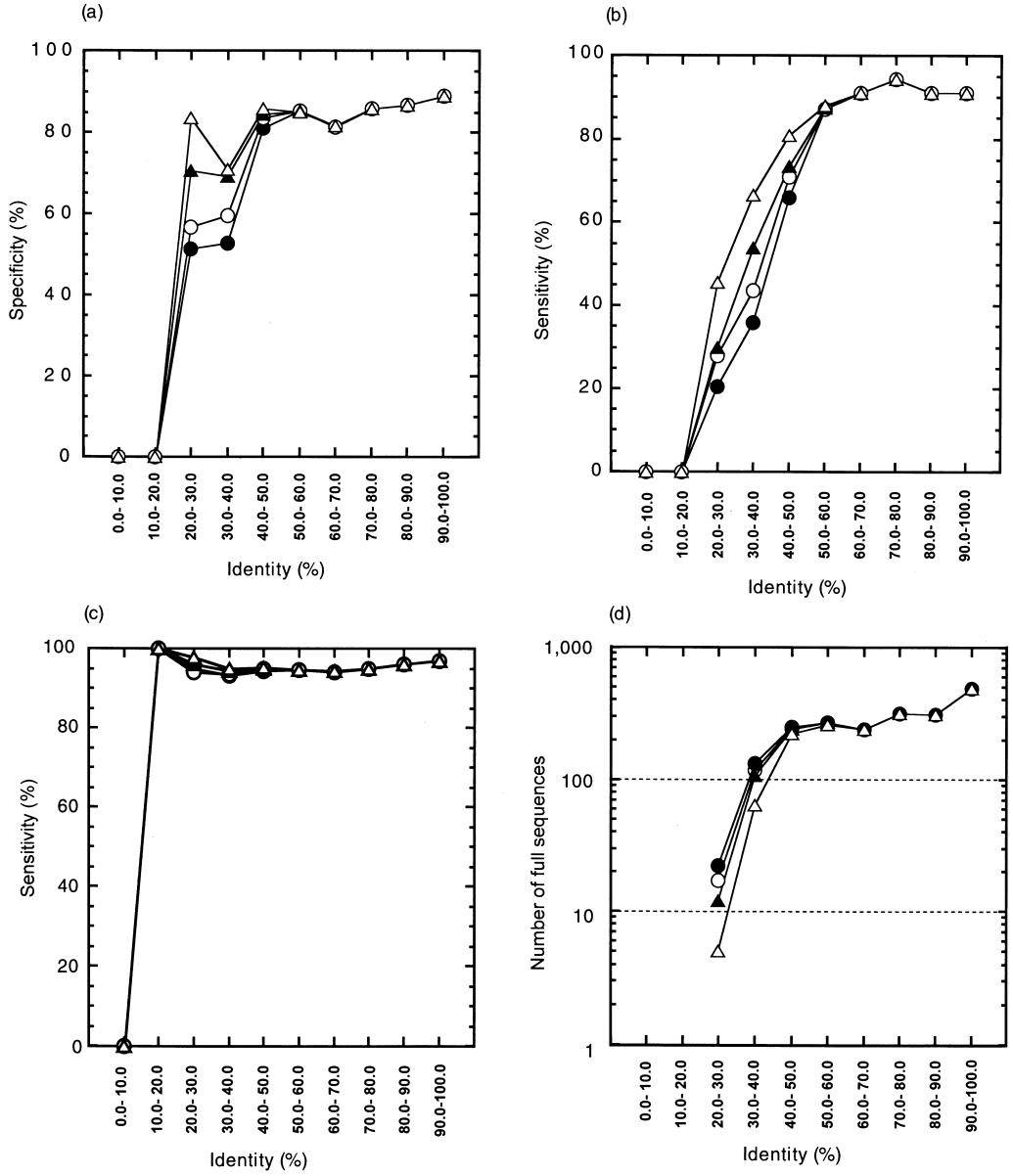

Fig. 6. Accuracy of the prediction by similarity in each identity subset. [Conditions: consensus length $\geqslant 50$ bases; E-values $\leqslant 1$ (●), $\leqslant 0.1$ (○), $\leqslant 0.01$ (▲), and 0.000 01 (△)]. (a) Specificity of positive prediction (the ratio of sequences truly predicted as full in sequences predicted as full). (b) Sensitivity of positive prediction (the ratio of sequences predicted as full in truly full sequences). (c) Sensitivity of negative prediction (the ratio of sequences truly predicted as not-full in not-full sequences). (d) Number of sequences truly predicted as full.

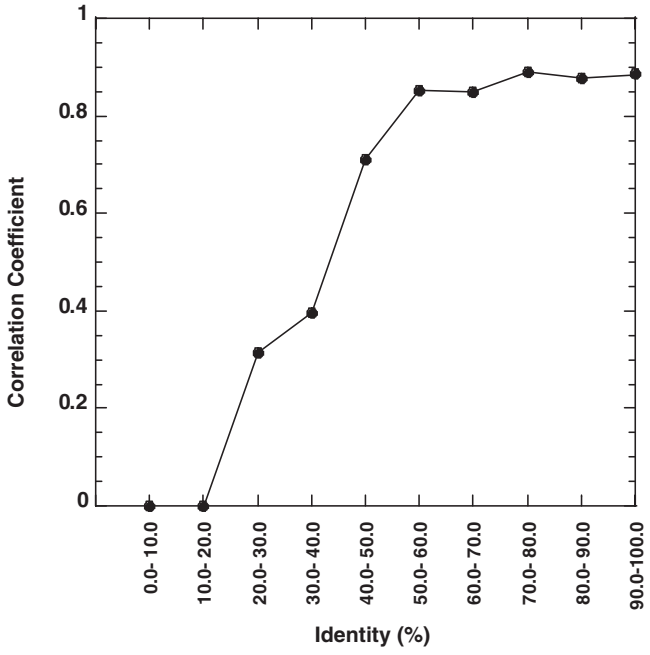

Fig. 7. Correlation coefficient of the positive prediction by similarity in each identity subset. [Conditions: consensus length $\geqslant$50 bases ; E-values $\leqslant$1].

In Figure 6d, under the same conditions as the other figures, the number of sequences truly predicted as full in each identity range is shown. This figure shows that the number of sequences truly predicted at identity range of 20–40% is larger at higher E threshold. Therefore, if false predictions over the higher E threshold are removed by any other method, both the number and the proportion of truly predicted sequences can be maximized. To remove these false predictions, the ATGpr score can be used effectively, as explained in the next section.

Prediction of fullness of cDNA by using ATGpr score and similarity with other proteins

Fullness prediction by both similarity and ATGpr score was performed and its accuracy is evaluated as follows. The prediction space, which consists of identity and ATGpr score, is divided into disjoint subspaces. Under the conditions that consensus length $\geqslant$50 bases and E-value $\leqslant$1, a plain having two axes (identity and ATGpr score) is divided into one hundred subspaces: (identity (%), ATGpr score) = (0–10, 0–0.1), (0–10, 0.1–0.2), ł. . . (0–10, 0.9–1), . . . (90–100, 0–0.1), (90–100, 0.1–0.2), ł. . . (90–100, 0.9–1). The accuracy of the predictions in each subspace is shown in Figure 8. Specificity of positive prediction (Sp) over the identity range of 20–60% is shown in Figure 8a. And specificity of positive prediction over the identity range of 60–100% is shown in Figure 8b. These figures indicate that Sp is greater than 90% over the identity of 20–30% if ATGpr Score is greater than 0.6, Sp is greater than 80% over the identity of 30–40%, if ATGpr Score is greater than 0.3. In the prediction using only similarity, however, Sp is approximately 50% over the identity of 20–40% (as described in the former section). Therefore, by adding ATGpr score condition to identity condition, Sp increases from 50% to greater than 80% over the identity of 20–40%. This increase is considered due to the removal of false predictions by the combination of identity and ATGpr Score. Figures 8a and b also show that Sp isgreater than 80% over the identity of 40–50% if ATGpr Score is greater than 0.2 and that Sp is greater than 80% at identity greater than 50% if ATGpr Score is greater than 0.1. As a summary, any identity range greater than 20% can increase Sp greater than 80% by combining an appropriate condition of ATGpr Score.

The characteristics of the fullness-prediction accuracy are well represented by the contour lines on the two-dimensional prediction space (identity-ATGpr score) in Figure 9. Specificity of positive prediction defined in Figure 8 is represented by contour lines in Figure 9a. Specificity of negative prediction is represented by the contour lines in Figure 9b. Specificity is smoothed by linear interpolation as a function of identity and ATGpr score before representing the contour lines. These figures clearly show the prediction subspace where specificity is greater than 80%, which corresponds to Figure 8.

Discussion

In this paper we newly used sequence similarity information for predicting fullness of cDNA fragment sequences. So far the relationship between sequence similarity and protein structure similarity has been studied (Sander and Schneider, 1991). The evaluation results in our study show that the accuracy of fullness prediction is greater than 80% if sequence identity is greater than 40% and the accuracy decreases to 20% if sequence identity decreases to 20%. This finding resembles the fact that structure similarity is characterized when sequence similarity is greater than 30%. Moreover, we newly showed that the prediction accuracy can be greater than 80% even when sequence identity is from 20–40% by combining sequence similarity information with statistical information of sequences (ATGpr score). The combination of sequence similarity and statistical information might be effective in solving the structure prediction problem, in which such an approach has not been used.

In exon prediction of genome sequences, it is known to be effective to combine sequence similarity with proteins and statistical information (Xu *et al.*, 1997). In this case, however, similarity information has to be dealt with carefully when evaluating prediction accuracy; that is, if

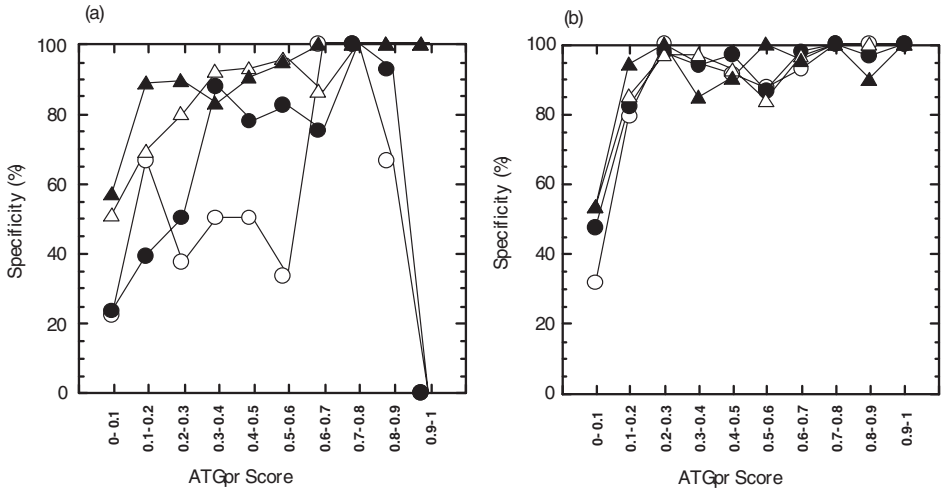

Fig. 8. Specificity of positive prediction by using ATGpr score and similarity in each identity-ATGpr score region when consensus length $\geqslant$50 bases and E-value $\leqslant$1. (a) Specificity of positive prediction is plotted over the identity subspaces from 20–60% (20–30% (○), 30–40% (●), 40–50% (△), 50–60% (▲)). (b) Specificity of positive prediction over the identity subspaces from 60–100% (60–70% (○), 70–80% (●), 80–90% (△), 90–100% (▲)).

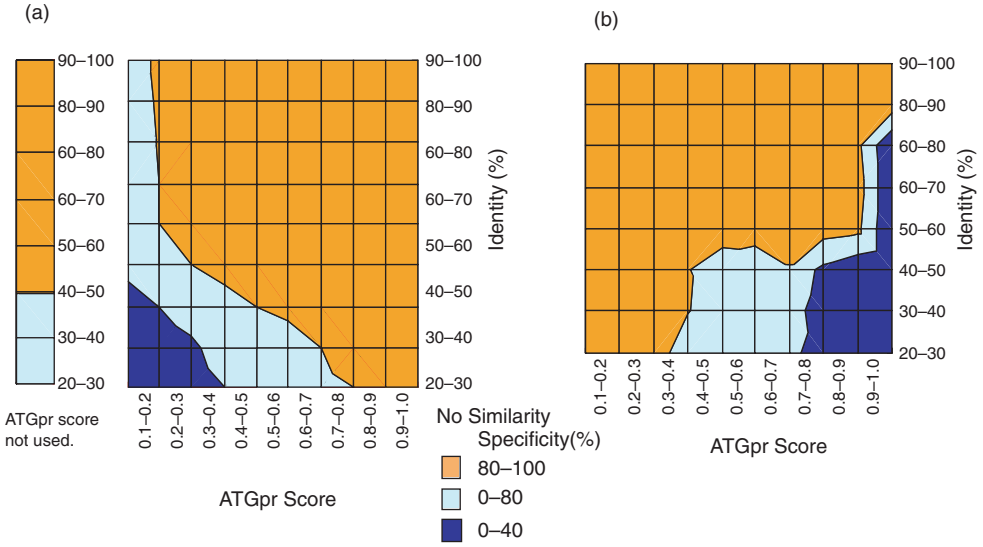

Fig. 9. Accuracy representation of fullness prediction by contour lines on two-dimensional identity-ATGpr score space. (a) Specificity of positive prediction. (b) Specificity of negative prediction. For both figures the subspaces are classified by shading. gray region: specificity $\geqslant$80%, white region: specificity $\geqslant$40% and specificity $<$80%, dark gray region: specificity $<$40%.

we predict using all similarity information greater than a threshold, the prediction accuracy of a gene having higher similarity with known protein sequences tends to be too high. Therefore, the prediction accuracy for an unknown gene might be lower than that for a known gene because an unknown gene has less similarity with known protein sequences than that for a known gene (unpublished). To solve this problem, similarity greater than a threshold was not used in our approach, but similarity was divided into several ranges and prediction accuracy in each range was examined. We can estimate the prediction accuracy for a given unknown gene by using this method.

Introducing similarity into discriminant analysis might be an effective method to combine the similarity with ATGpr score. The purpose of this paper is, however, to clarify

the effectiveness of the similarity itself and to know the complementary relation between the similarity and ATGpr score relevant to the prediction of the fullness of a cDNA sequence. Therefore, we introduced a two-dimensional space consisting of the similarity and ATGpr score and we evaluated the prediction in each disjoint subspace of this two-dimensional space. This two-dimensional expression can display the contribution of the similarity and ATGpr score to the fullness prediction intuitively. We therefore would like to place it as the next target to deal with the fullness prediction with an extended discriminant analysis that combines similarity and ATGpr score.

Acknowledgements

We would like to thank Y.Nakamura and T.Nagai of the Helix Research Institute for constructing a system in which we can use informatics tools.

References

Hillier,L.D. *et al.* (1996) Generation and analysis of 280 000 human expressed sequence tags. *Genome Res.*, **6**, 807–828.

Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.

Salamov,A.A., Nishikawa,T. and Swindells,M.B. (1998) Assessing protein coding region integrity in cDNA sequencing projects. *Bioinformatics*, **14**, 384–390.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet.*, **9**, 56–68.

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.

Xu,Y., Mural,R.J. and Uberbacher,E.C. (1997) Inferring gene structures in genomeic sequences using pattern recognition and expressed sequence tags. In *Proceedings of 5th International Conference on Intelligent Systems for Molecular Biology* pp. 344–353.