ESTFastAnnotator-High Throughput Protein Domain Prediction on EST by Utilizing InterPro

Abstract

ESTFastAnnotator provides a fast, systematic, and automatic EST protein domain annotation method, which overcomes the limitations of EST sequence carried in nature, and eliminates the impact of E value cutting threshold. The precision and recall rate of ESTFastAnnotator keeps over 91% under different E values, and the total running time is at least 13 times lower than InterProScan (which is mostly composed by sequence-motif search tools). ESTFastAnnotator can explore more function on short EST sequences than sequence-motif search tools by 84.2% of correct hit rate, as InterProScan is 34.3% on the same EST.

Keywords: EST, protein, domain, prediction.

1. Introduction

Since the first disclosure of Human Genome project in 1984 [8], the speed of gene discovery research has been changed dramatically. The majority of human genes; once seem to remain unknown for the next decade, have now fully revealed their genetic code with the completion of human genome [7]. All along, gene discovery has been developed in two different aspects [1]. Proponents of cDNA (EST: Expressed Sequence Tag) sequencing claimed that the coding sequences, only 3% of the DNA, have represent the majority of the information content, thus cDNA should take procedure over genomic sequencing [1]. On the other hand, proponents of genomic sequencing claimed that finding every mRNA expressed in all tissues, cell types, and development stages would be difficult, and pointed out that much valuable information from intronic and intergenic regions, including control and regulatory sequences, would be lost by cDNA sequencing [14]. Although the incomplete and error nature of EST occurs, cDNA sequencing still becomes the most powerful and efficient method in gene discovery [2].

The ESTs in public databases have been dramatic grown up since the original description of ESTs published in 1991 [1]. In mid 1995, the number of ESTs in GenBank surpassed the number of non-EST records [2]; as of June 2000, 62% of the sequences in GenBank are comprised of 4.6 million EST records. Large EST projects have accumulated over 11.3 million un-annotated files from more than 5318 cDNA libraries until January 2002 [5]. As of that, a rapid, large-scale, and automatic EST annotation work flow is essential in genome research [17]. Most commercially available softwares (e.g. Pfam [16], PROSITE [15], PRINTS [4], SMART [12], and TIGRFAM [9], now integrated into InterProScan [18]) have limited use in processing large-scale sequence data [17]. Although the software, such as BLAST [3], is designed to process large-scale DNA sequences and has been applied in most EST annotation workflows, large amount of user manipulation of cutting threshold selection and investigation is still in great demand. Besides, most EST sequences are short and low quality sequences, which make them rejected by sequence-motif methods easily [18]. Thus, exploring much information on these EST sequences can be a great assistance in novel gene discovery.

In this paper, we provide a novel, fast, and automatic EST protein domain annotation method, which still keeps high precision and under various cutting thresholds. This is achieved by fast sequence alignment and protein clustering under loose cutting threshold to get the overview of possible protein hits. In the main concept of this work, the protein domain of input EST is annotated by considering the hit protein list entirely, not only the top protein hits. All proteins are clustered according to sequence similarities and structure similarities. The quality of annotation mostly depends on the refinement process of protein clusters. However, the refinement process of protein clusters is actually a two-class problem. In this work, a SVM classifier, which seeks an optimized solution and avoids over-fitting, is introduced to perform protein cluster refinements. We attempt to discover more information than traditional sequence-motif method by annotating EST through entire protein clusters, especially for short EST sequences.

2. Method

The whole process of ESTFastAnnotator is summarized in Figure 1. The automatic alignment is performed by BLAST [3] against SwissProt [6] under loose threshold to enlarge the protein search scope. After that, the features of each protein sequence are extracted according to InterPro mapping table (which
will be explained in next section). All proteins are clustered by the features of sequences. From the
features of each protein cluster, the SVM classifier can make decision and pick up possible protein clusters.

The extracted features of protein clusters are divided into three parts: supporting scores, similarity scores and structure relationships. The supporting score are used to measure the support degree of protein clusters between protein voters and InterPro voters. The possibility of a protein cluster gets higher when it is supported by more protein voters and InterPro voters. In addition to supporting score, the similarity scores are used to measure how significant a protein cluster is. The similarity scores can be used to filter the protein clusters with high votes but are actually supported by dissimilar protein voters. Besides supporting scores and similarity scores, the structure relationship between protein clusters is also an important feature. The structure relationship is not completely coincident with supporting score or similarity score, especially for remote proteins. The possibilities of protein clusters with low supporting scores or similarity scores will be raised when they have structure relationships with protein clusters of high supporting scores or similarity scores. These features are listed in Table 1.

2.1 InterPro

InterPro [11] is an integrated documentation resource for protein families, domains, and sites. InterPro combines a number of databases (e.g. Pfam [16], PROSITE [15], PRINTS [4], SMART [12], and TIGRFAM [9]) that use different methodologies and a varying degree of biological information on well-characterized proteins to derive protein entries. Each entry in InterPro represents a group of proteins which have similar sequence and similar structure, including families, domains, sites, repeats…etc. Thus, the hit proteins of our system can be clustered by simply search against the InterPro mapping table.

InterProScan [18] is a tool that combines different protein signature recognition methods into one resource. Searching InterPro through InterProScan is equal to perform pattern recognition in all member databases simultaneously. In this work, InterProScan is used to produce the desired output. The running time and EST hit rate of our work is compared with InterProScan.

2.2 Feature Extraction

The extracted features of protein clusters are divided into three parts: supporting scores, similarity scores and structure relationships. The supporting score are used to measure the support degree of protein clusters between protein voters and InterPro voters. The possibility of a protein cluster gets higher when it is supported by more protein voters and InterPro voters. In addition to supporting score, the similarity scores are used to measure how significant a protein cluster is. The similarity scores can be used to filter the protein clusters with high votes but are actually supported by dissimilar protein voters. Besides supporting scores and similarity scores, the structure relationship between protein clusters is also an important feature. The structure relationship is not completely coincident with supporting score or similarity score, especially for remote proteins. The possibilities of protein clusters with low supporting scores or similarity scores will be raised when they have structure relationships with protein clusters of high supporting scores or similarity scores. These features are listed in Table 1.
Table 1  Features of Protein Cluster

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>ID of protein clusters</td>
</tr>
<tr>
<td>Supporting Score</td>
<td></td>
</tr>
<tr>
<td>HitPercent_voter</td>
<td>The affirmative voting percentage of protein voters</td>
</tr>
<tr>
<td>HitPercent_node</td>
<td>The percentage of protein voters in InterPro node</td>
</tr>
<tr>
<td>Similarity Score</td>
<td></td>
</tr>
<tr>
<td>MaxBitScoreRatio</td>
<td>The max bitscore of proteins associated to the protein cluster</td>
</tr>
<tr>
<td>SumBitScoreRatio</td>
<td>The sum bitscore of proteins associated to the protein cluster</td>
</tr>
<tr>
<td>Structure Relationship</td>
<td></td>
</tr>
<tr>
<td>IsRelationExists</td>
<td>Is there any parent-child relation of this protein cluster? (Y/N)</td>
</tr>
<tr>
<td>IsRelationWinESTExists</td>
<td>Is there any parent-child relation between this protein cluster and other protein clusters within same EST? (Y/N)</td>
</tr>
<tr>
<td>Cluster_type</td>
<td>Family, Domain, Site….etc.</td>
</tr>
<tr>
<td>Cluster_level</td>
<td>The level of protein cluster in InterPro Parent-Child tree</td>
</tr>
</tbody>
</table>

Figure 2 shows the ROC curves of supporting score (HitPercent_voter) and similarity score (MaxBitScoreRatio) under different E values. From the graph it can be seen that supporting score has excellent separating ability of two classes. The separating ability of supporting score is not effected by E values. This is because the physical meaning of supporting score is not related with E values. The true effect of E value is on similarity score. From (b) we can see the separating ability decreases as E value. Similarity score is used to filter the impossible protein cluster, which is useful under loose threshold (high E value.) Actually, the best operating condition of ESTFastAnnotator is under loose threshold, which makes the protein search scope as large as possible.

2.3 Dataset

The training and testing dataset is mainly extracted from TIGR Rice Gene Indices [13]. The version of TIGR Rice Gene Indices is Release 14.1 released on January 16, 2004. Each gene in TIGR is reconstructed by EST and gene fragments. Hence, the training set is all EST sequences of the genes. The system is evaluated by 10 folds cross validation.

2.4 Desired Output

EST is a fragment of gene. Most EST sequences are shorter (fuzzier) than genes, means they may be similar to more protein clusters than genes. On the other hand, the protein clusters associated with each EST sequence, if can be found, should be contained in the protein clusters associated by the whole gene. Upon all protein clusters of the EST, only the clusters appear in the gene can be viewed as correct protein clusters. EST sequence should only contain partial protein clusters of the gene; only the EST of full length will have exactly the same protein clusters as gene. The relationship of gene, EST and protein clusters is illustrated in Figure 3. The protein clusters of gene can be obtained by InterProScan, which are protein clusters A, B and C in Figure 3.
The EST on the left hand should contain partial protein cluster of the gene, which are the desired output.

![Figure 3: Protein clusters associated with gene and EST sequences.](image)

3. Result

Figure 4 shows the precision and recall rates under different E values (cutting thresholds). The plot shows the precision rate keeps over 90% and recall rate keeps over 91% under different E values. The average precision is 95% and average recall is 93.4%. Different choice of E values will not have strong influence on this system.

![Figure 4: Precision and recall rates under different cutting thresholds.](image)

From this result we can make sure that ESTFastAnnotator can select possible protein cluster correctly.

Figure 5 represents the running time and correct EST hit rate compared with InterProScan. The running time of our method is much shorter than InterProScan. This is because most member tools of InterProScan are HMM (Hidden Markov’s model) search. The time needed to calculate all paths of HMM model is relatively long. The speed of ESTFastAnnotator is enhanced by rapid sequence alignment tool BLAST and SVM. To be noticed, the running time of both increases as the length of EST extends, but the increasing rate of ESTFastAnnotator is lower than InterProScan.

The correct EST hit rate of ESTFastAnnotator is higher than InterProScan; especially for those EST sequences which are 100-300 bp in length. For EST sequences over 500 bp, the correct hit rate of ESTFastAnnotator and InterProScan will close to each other gradually. As for ESTs over 700bp, the correct hit rate of ESTFastAnnotator and InterProScan will both be 100%. However, ESTFastAnnotator has a hit rate of 92% in average, which is higher than 79% of InterProScan. ESTFastAnnotator can explore more information than sequence-motif method, especially on short length ESTs.
4. Discussion

The results of Figure 5 display the superiority of ESTFastAnnotator over InterProScan on EST annotation. The time of ESTFastAnnotator needed for all lengths of ESTs is much shorter than InterProScan. The kernel of ESTFastAnnotator is BLAST and SVM classifier, both has good performance in speed and sensitivity, makes time complexity decrease a lot. As for InterProScan, the bottleneck lies in hidden Markov model search (e.g., Pfam, SMART, and TIGRFAM).

ESTFastAnnotator has 92% of hit rate in average, which is higher than 79% of InterProScan. The gap between ESTFastAnnotator and InterProScan is large for short length EST sequences (84.2% for ESTFastAnnotator and 34.3% for InterProScan at length 100-300 bp). However, the gap between ESTFastAnnotator and InterProScan decreases as the length of EST sequence increases. For long or full length EST sequences, both the hit rate of ESTFastAnnotator and InterProScan reaches 100%.

Figure 6 shows the limitations of InterProScan on EST annotation. The profiles of InterPro are actually extracted from proteins, which are usually longer than translated ESTs. Short ESTs will be rejected by these protein profiles easily. On the contrary, ESTFastAnnotator considers the hit proteins entirely before making conclusions, the informations of short EST will not be lost.

But for InterProScan, if the profile of short EST can be found, the precision and confidence will be higher than ESTFastAnnotator. Meanwhile, the domains predicted by ESTFastAnnotator are usually multiple, and the confidence of each domain is less than the domain predicted by InterProScan.

Protein A ....AAGHVNIAEAVQPLNHR
Protein B ....BAGMVNMAEAQPLNHR
Protein C ....AGHVNIINEAMQQLGHR
EST Translated QPLNHR
profile VNIM---Q---L- - R

Figure 6 Limitation of sequence-motif method

5. Conclusion

ESTFastAnnotator is not intended to fully replace existing sequence-motif based or homology search, but to provide alternative data/patterns from the EST sequences to go beyond traditional approaches. ESTFastAnnotator proposes a fast, systematic and automatic function annotation method, which maintains high precision and recall under various cutting thresholds. It is especially superior in exploring functions of short EST sequences, which cannot be annotated easily by current sequence-motif methods.

The best role for ESTFastAnnotator is to explore new information of novel ESTs and novel genes. ESTFastAnnotator is superior in exploring protein domains of EST sequences which have difficulty to be annotated by current methods. For those EST sequences which cannot find any hit by InterProScan, ESTFastAnnotator can be a good supporting search tool.

6. Future Work

In this work, the architecture of refinement process is one single SVM classifier. The overall performance can be further enhanced by hierarchical protein cluster selection.

References


