A Proposed Paradigm for Expressed Sequence Tags Data Format – An Application of Hidden Markov Models

Yen-I Chiang1,2,* and Guan-I Wu1

1Department of Information Management, Chang Gung University, Taoyuan, Taiwan, R.O.C.
2Bioinformatics Center, Chang Gung University, Taoyuan, Taiwan, R.O.C.

Abstract

In the era of post-Human Genome Project, researches have shifted the emphasis from the mapping of human genomic to the discovery of correlation between genetic markers and clinical phenotypes, where finding effective treatment against disease are becoming crucial and applicable goals. The Expressed Sequence Tags (ESTs) data plays an important role in the completion of the Human Genome Sequencing and is widely used for gene discovery, polymorphism analysis, expression studies, and gene prediction. However, due to the chemical properties and manufacturing processes, ESTs data might contain errors, which might mislead Bioinformatics researchers that attempt to use EST-libraries to identify Single Nucleotide Polymorphisms (SNPs). Therefore this study proposes a paradigm for EST data, where users might better address this issue and use them to correctly identify SNPs.

Key Words: Expressed Sequence Tags, Hidden Markov Models, Base-Calling, Electropherogram

1. Introduction

The study of Human Genomic has come a long way, since the inception of the Human Genome Project in 1986. In the post-Human Genome Project era, the main research focus has shifted from the Human Genome Sequencing into applications of genetic technology [1]. In the short span of no more than 20 years, the data generated by various research centers is accumulating at an incredible speed, and databases have increased hundred-fold [2]. With such large amount of data been accumulating and generating even now, the question of how to use those data efficiently and accurately has become a major issue. Furthermore, the research emphases have shifted from completing the map for human genome to the discovery of genetic markers and corresponding diseases, where finding effective treatments against diseases are becoming crucial and applicable goals. Therefore, more attention should be paid to finer details of the data in the existing databases [3–5]. In fact, the completion or near-completion of the map for human genomic has created opportunities for studying numerous aspects of gene function at the genomic level that was impossible prior to the completion of the human gene catalog; where the problem is considered to be too complex to be addressed.

Expressed Sequence Tags (ESTs) database is one of such class of databases [6–9], where enormous amount of wealth are waiting to be mined. ESTs are short nucleotide sequences with average length ranging from 400 bases to 600 bases and are obtained from single-pass sequencing of cDNA clones that are sequenced from 5’ to 3’. ESTs are widely used throughout the genomics and molecular biology communities for gene discovery, mapping, polymorphism analysis, expression studies, gene prediction and are instrumental to the rapid completion of the Human Genome Sequence Map. Many commercially established and privately funded centers maintain various ESTs libraries with sources of many of those
ESTs data from in-house collections.

Numerous computational sounded methods have been developed for Bioinformatics. However some of those methods often yielded results being far below the expectation and rendered those methods impractical to be used in clinical analysis. The fault might cause by the quality of the input data and error-containing ESTs data. One of the reasons is the technical limitations on sequencing methods resulting in sequencing errors in EST raw data even though ESTs contain wealth of information on DNA transcription. It creates a rift between the bioinformatics and molecular biology communities where results using high throughput methods are used with skepticism and for exploratory purposes only [5,10,11]. Increasingly, scholars are starting to realize this and are actively in addressing the issues [12–14].

The erroneous input sources due to the technical limitation have created the undesirable result mentioned previously, thus there is a need to implement an automated scheme that can be used to accurately identify and cluster transcript sequences. This study proposed to implement an automated and high-throughput scheme to filter out and mark out the erroneous regions of each EST sequence using Hidden Markov Models [15–19], so that EST sequences maybe accurately cluster together and give users more reliable information when locating SNPs [20,21]. The purpose of this paper is not to replace the existing base-calling programs such as Phred or Phrap [22] but to facilitate them, and in the meantime to point out the hidden errors embedded in the EST sequences data during the process.

The paper is organized as follow: Section 2 gives the Biological background information of this study. Section 3 states and formulates the problem. Section 4 presents a scheme that addresses the problem of the EST data set. Section 5 contains a summary of the computational result generated from the scheme and Section 6 gives a conclusion.

2. Background

2.1 Single Nucleotide Polymorphisms

It is obvious to us that people are different. For example, people have distinct physiques and abilities, and some people are more susceptible to certain diseases than others. Many of these kinds of differences among people can be explained by their genetic variations – alterations in the DNA sequences that regulate the building of important proteins [23–25]. Sometimes the alterations involve a single base pair (the smallest building block of DNA) and are shared by many people. Such single base pair differences are called “single nucleotide polymorphisms” or SNPs for short. However, the majority of the SNPs do not produce physical changes in people with affected DNA.

Analysis of human genetic variation can shed light on the problem of the genetic basis of complex disorders [26–29], hence there are so many genetic scientists eagerly to identify as many SNPs as they can. SNPs, the most common DNA variants in humans, represent a valuable resource for the genetic analysis of cancer and other illnesses. SNPs that do not themselves change protein expression and cause disease may be close to the chromosome of deleterious mutations. Using this proximity, SNPs may be shared among groups of people with harmful but unknown mutations and can be used as markers for them. Such markers can be used to locate mutations and accelerate the processes of finding therapeutic drugs. Furthermore, analyzing shifts in SNPs among different groups of people will help population geneticists to trace the evolution of the human race down through the millennia and to unravel the connections between widely dispersed ethnic groups and races. Complete enumeration of non-synonymous SNPs in the candidate genes will enable further association studies on panels of affected and unaffected individuals; ESTs are being used extensively to identify SNPs [30].

2.2 Expressed Sequence Tags

As reported in various literature, a rapidly growing area of genome research is the analysis of ESTs [1]. ESTs are short partial DNA sequences, usually of length 150 bp to 1000 bp, generated from the 3’ and 5’ ends of randomly selected cDNA clones from specific tissues. The resulting ESTs reflect the level and the complexity of gene expression in the sampled tissue. ESTs can be used to rapidly identify expressed genes because they are usually unique to the cDNA from which they are derived and are corresponding to a specific gene in the genome. The purpose of EST sequencing is to scan rapidly for expressed genes and to provide a tag for each gene. Due to the fast sequencing process, the large number of sequ-
enced ESTs has been generating and been storing into databases. Thus EST databases contain a potential wealth of valuable information about expressed genes [31]. The fragmented nature and vast quantity of EST data pose an obstacle to harvest the full potential from this data source, where the fragmented, error-prone EST data and the known gene sequence data can be consolidated and placed into the correct cluster. Several efforts are under way to condense single-read ESTs and full-length transcript data on a large scale by means of clustering or assembly, where they want to store all data concerning a single gene into a single cluster or a specified class of clusters (i.e. the sequence contains in a cluster if and only if they represent the same gene).

The problems with collections of EST sequences that make them difficult to use are that these data are often highly redundant, incomplete, and low in quality when been clustered together. Clustering and assembly methods can alleviate the problem of redundancy and improve the quality and length of consensus sequence; however, these techniques tend to lead to different views on the data, depending on the exact method used and criteria/interpretations used by end users. To derive information from these large and diverse data is a nontrivial task. Some computations and algorithms are needed, where cluster analysis in which sequence similarity and possibly other criteria are used to form the clusters or the index classes.

3. Motivation and Formalization of the Problem

The task of removing redundancy and accurately clustering the ESTs into the correct gene clusters is a challenging problem not only because of the imposing size of the EST databases but also because of intrinsic difficulties arising from low sequence quality, highly similar (but distinct) gene family members, problematic cDNA clones, retained introns and alternatively spliced transcripts, incomplete gene coverage, and other limitations. Here a brief general description of a scheme that is used to cluster EST data set is given:

1. ESTs are put into distinct clusters such that each cluster represents a distinct gene, including all alternative transcript isoforms derived from the same gene. Various tools such as Blast algorithm and UniGene index can be used and can be found at http://ncbi.nlm.nih.gov/UniGene/.

2. Each EST cluster is deemed to represent a distinct mRNA transcript. In particular, alternative transcript isoforms are represented by distinct EST clusters. This strategy is implicit in DNA assembly tools such as Phrap (http://www.phrap.org), TIGR Assembler (http://www.tigr.org/tdb/tgi.shtml) then ESTs are categorized by their RNA source and are subsequently clustered separately for each source sample.

Procedures described above need subjective interpretation form experimenters or analyzers, because data is obtained from a restricted technique. The primary evidence for ESTs representing the same gene or transcript is derived from significant sequence similarity in a region of sequence overlap. Using fragment assembly algorithms, ESTs can be clustered to form a consensus sequence corresponding to each putative transcript; significant sequence similarity must be able to incorporate mismatches and insertions/deletions (dynamic programming algorithms for alignment can be used). However, as previously mentioned, the techniques used to generate EST data are not 100% perfect, where there are regions of low sequence quality [32,33]. A given electropherogram of an EST sequence contains signals of vary magnitude, where a typical good signal will show signals for “A”, “T”, “C”, and “G” and should be contained in certain range of signal values. Peaks are sharper for the first 500 bases and resolutions of peaks drops off from that point, but are still readable up to 750 bases. Here is a summary of the possible erroneous signals found in an electropherogram of an EST sequence:

1. The initial transcribed segment of an EST sequence: it is difficult to hybridize onto the template due to limitations of the instrument used, the chemical structure of the primer chosen and reaction of the enzyme used.

2. The transcribe segment near the termination sector of an EST sequence: there is a limitation to the length of hybridization, where the process becomes unstable when it passes a certain point. This is due to the fact that enzyme reaction becomes unstable and can not sustain for so long and wrong bases are being used for transcription. Furthermore, the quality of the sample clone can affect this length drastically.

3. Lengths of erroneous segments vary from 30 bp to
100 bp, depending on the skill of the experimenters and the quality of the clones. The middle portion of EST seems to be the most accurate section.

4. Erroneous segments might be embedded among the section of sharp, well defined signals due to interactions between the 4 chemical solutions used in the process of peak characterization in electropherograms.

Sequences at the tail-ends of an EST sequence played important roles in the assembling of any genome sequence map, where EST sequences with similar overlaps are aligned and clustered into a single transcript. Hence any inclusion of an erroneous segment might result in the misrepresentation of a gene. Conversely, any exclusion of a useful segment will result in the shortening of the transcript, where improper transcript was recognized and rendered the transcript unusable. As a consequence, valuable information might be lost. Furthermore, the inclusion of quality doubtful base calling will result in the false identification of SNPs.

Hence the construction of an automated program to solve this problem is entailed. A refined or accurate alignment for clustering is computationally too intensive. Hence, alternative means of approximate overlap detection [34,35] can be used for fast identification of sequences with potential for good quality overlap (with a preset threshold and a given acceptable error rate). Usually there will be thousands of sequences left in the cluster still, and then a refined algorithm will be used to accurately cluster those EST sequences.

4. Method and System Overview

Over time, methods that automate DNA sequencing to generate the sequencing trace data (electropherograms) have been refining and improving, and much knowledge and experiences have been accumulating from the sequencing chemistry, where one might expect to obtain results in nearly even peak intensity patterns in the sequencing trace. Detail explanations can be found in the Chemistry Guide by Applied Biosystems Inc. [36]. However, in practice this is rarely the case, the trace data usually contains unsatisfactory sections for the entire length, where uneven patterns were observed. The effect of uneven peak intensity patterns might decrease the accuracy of base-calling and make detecting differential sequences more difficult [36–38]. Various observed common peak patterns are recorded in [39] such as repeated weak “G” signals after “C” signals, stair-step pattern of repeated “G” signals and weak “G” signals after “A” signals etc.

Usually, some criteria or schemes will be applied to filter out the error-containing sectors. Current methods that follow the threshold types of approach are sequence-independent with too much or too little of sequences being removed, resulting in mis-clustering or diminishing the ability to cluster sequences; hence they are highly unsuitable for large volume clustering. In this study, a decision rules based approaches will be implemented to filter out the low sequence quality end-sections of each EST, using the respective electropherogram, because the shapes of those erroneous signals are very distinctive.

A well-developed and acceptable tool, such as PHRED, will be used to extract reading from the outputs of DNA sequencing equipments and perform the initial base-calling. In addition, the hidden Markov Models is used to identify the erroneous signals embedded in each putative sequence, where any highly devious base-calling will be marked and indexed [40,41].

The current version of this system is able to observe seven major patterns as indicated in Table 1. These patterns can be described as erroneous sections transcribed on the electropherogram data from the CE instrument (called the trace files). A fully Bayesian approach of hidden Markov Models will be constructed for each pattern and then some of the trace files will be used to test the efficiency and the responds of this model. Depending on the outcomes of the evaluation, other alternative hidden Markov Models such as the stochastic-context-free grammars approach might be considered for future study.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern 1</td>
<td>strong T signal, preceded by weak G signal</td>
</tr>
<tr>
<td>Pattern 2</td>
<td>weak G signal, preceded by strong T signal</td>
</tr>
<tr>
<td>Pattern 3</td>
<td>consecutive appearances of uniform strength As after signal C</td>
</tr>
<tr>
<td>Pattern 4</td>
<td>strong C signal, preceded by weak T signal</td>
</tr>
<tr>
<td>Pattern 5</td>
<td>weak G signal, preceded by signal C</td>
</tr>
<tr>
<td>Pattern 6</td>
<td>weak G signal, preceded by signal A</td>
</tr>
<tr>
<td>Pattern 7</td>
<td>Strong A signal, preceded by weak G signal</td>
</tr>
</tbody>
</table>
Classical algorithms such as Baum-Welch algorithm, Viterbi algorithm even some of the heuristic method can be incorporated to assist in the processes of estimating the parameters.

Analogous to various works done previously [42, 43], this study proposed to incorporate information of uneven peak patterns into the base calling procedure, where consequent analysis and procedures can benefit and deal with those erroneous sequences. A Hidden Markov model (abbreviate as HMM) that is able to distinguish the common uneven peak patterns is implemented (in this study, only seven different patterns were incorporated), and a Markov chain is modeled for each pattern to be recognized. For clarification, a state diagram in Figure 1(a) is used to illustrate this. Those seven Markov chains models are then combined into a single model that is capable of recognizing most of the uneven patterns (see illustration in Figure 1(b)). Next, a set of state transition probabilities, state emission probabilities and begin state probabilities were assigned for this model, and then use the Viterbi training algorithm to decode the obviation sequences [40, 44].

The HMM application system is implemented in Java and incorporated with Biojava API, containing six main modules: Trace file reader, Preprocessing, Data transformation, Parameter adjusting, Optimization and Electropherogram drawing. Figure 2 shows the schematic diagram of this system. Below is a description of each module:

**The trace file reader module**

This module reads the electropherogram data (raw sequence trace files from the automatic sequencing machine) saved in .ABI or .SCF files directly.

**The preprocessing module**

This module is used to segment the pattern of interest from the background by removing noise and normalizing the pattern etc. [33]. As a matter of fact, both ends of the raw sequence trace values (with varying lengths) contain low quality signal that os needed to be trimmed from the input. To speed up the process, the section from the begin position to next 200 bases, and the section starting from the 450 base to the end of sequence were scanned, and then followed by the trimming of low quality sequence segments from the input as indicated in Figure 3.

**The data transformation module**

The module extracts the appropriate features and

![Figure 1](image-url)
reading for each base, which will be input into the Viterbi procedure in the Parameter adjusting module. Based on the mean intensity and corresponding standard deviation of each set of peak trace values, each nucleotide’s signal value will have been classified into five classes of peak heights. To describe this module more precisely, the process is formulated as follow: let \( S \) be a trimmed input sequence that starts from position \( m \) and ends at position \( n \) (where \( m < n \)), then the base-calling process will convert \( S \) into a string \( s_m \ s_{m+1} \ s_{m+2} \ldots \ s_n \), where \( s_i \in \{A, T, C, G\} \) for \( i = m, m+1, \ldots, n \). Furthermore, the maximum signal intensity for any base-calling \( s \) will be denoted as \( \text{Peak}(s) \). The transformation procedure expects to embed the relative signal intensity of each base in \( S \) with respect to the sequence as a whole, therefore the mean peak intensity and the corresponding standard deviation of each base-calling signal were calculated using Eqs. (1) and (2):

\[
\bar{x} = \frac{\sum_{i=m}^{n} \text{Peak}(s_i)}{n-b+1}
\]

\[
\sigma = \sqrt{\frac{\sum_{i=m}^{n} (\text{Peak}(s_i) - \bar{x})^2}{n-b+1}}
\]

Then the signals were categorized into 5 distinct classes with respect to those two values. Based on the category of the signals, the HMM interprets signals’ relation to each other. Eqs. (3)~(7) are the criteria used for the signal categorization.

\[
\text{category } 2 = \{ s \mid \text{Peak}(s) > \bar{x} + 2\sigma \} 
\]

\[
\text{category } 1 = \{ s \mid \bar{x} + \sigma < \text{Peak}(s) \leq \bar{x} + 2\sigma \} 
\]

\[
\text{category } 0 = \{ s \mid \bar{x} - \sigma \leq \text{Peak}(s) \leq \bar{x} + \sigma \} 
\]

\[
\text{category } -1 = \{ s \mid \bar{x} - 2\sigma \leq \text{Peak}(s) < \bar{x} - \sigma \} 
\]

\[
\text{category } -2 = \{ s \mid \text{Peak}(s) < \bar{x} - 2\sigma \} 
\]
For example, the input sequence C C G G A G A C G A G A C C A may now be transformed into C⁻¹ C⁰ G¹ G² A² G⁻² A⁻¹ G⁻¹ A² C⁻¹ C⁻¹ A⁻², where each base-calling signal is indexed into the respective class (see Figure 4 for an illustration).

The parameters adjustment module

This module adopts the Viterbi Path Counting algorithm [45,46] for training parameters. The reason that Viterbi Path Counting algorithm is preferred over the typically used Baum-Welch algorithm in HMM, is because the main goal here is to recognize predefined patterns and adjust the parameters’ values accordingly.

The optimization module

This module uses the Viterbi algorithm to identify an optimal chain that can best describe the input and also to locate the segments of uneven peak patterns. Hence this module is in charge of optimization and prediction.

The electropherogram drawing module

This module controls the graphical outputs for this system as illustrated in Figure 5(a), (b) and (c). Users may decide the format of outputs that will suit their purposes most.

The resultant sequences will then be clustered using a standard multi-alignment algorithm. Transcripts of dif-
ferent individuals clustered from the resultant EST sequences of this algorithm can then be used to identify SNPs with a higher degree of confidence.

5. Dataset and Implementation Results

This work is performed under the Windows XP/Windows 2003 operating system, on a Pentium 4 1 GHz system with 256 MB of main memory and 1 GB of disk space. Data set was obtained from the Bioinformatics center at Chang Gung University, from the newly constructed EST library for Trichomonas vaginalis. Total of 80 raw sequences have been used to train and test for this system (however there are ten thousand folds of trace data in this library).

The statistics for result from the Preprocessing module on the unreliable signals trimmed away from both ends of the input sequences are summarized in the Table 2. Compared with the result obtained from an expert’s manual inspection of those 80 raw sequences, the system’s output result is closely corresponding with the result of the manual output, except for a few sequences. Those faulty sequences, which were not properly trimmed, are mostly caused by other base-calling errors and extremely rare, hence the preprocessing model is performing well. Based the result of false positive and the false negative shown in Table 2, the accuracy this system is satisfactory.

The performance of the HMM modules is then being evaluated, both for the time complexity and the correctness of this algorithm. In terms of the time complexity, with a proper set of parameters found through a training process of those modules, the run time used to process input data is extremely short. The correctness of this algorithm is being evaluated, base on the ten cross validation method, where raw sequences were randomly split into two sets, with one set being allocated for training and the other for testing. The resulting statistic is summarized in Figure 6, where a graph illustrates the average occurrence frequencies of each of the seven uneven peak patterns in every input sequence. As the graph illustrated, pattern seven, the occurrence of elevated intensity signals in As after appearance of weak G signal, is observed most frequently; pattern two, the occurrence of descending signal in G, preceding by strong signal T, is not observed by the models. After a manual inspection of randomly selected inputs, the result seems to be promising; the predicted sequences matched closely with the actual sequences’ signals. Lastly, the accuracy of this system is evaluated by a deterministic method for the identification of suspected segments of each input sequence and then all 80 resultant input sequences were manually verified by an experimental personnel. Next, each input sequence is segmented into imaginary units of four consecutive bases and the rate of false negative and false positive of each sequence for each pattern were evaluated (the only pattern that was unsuitable to evaluate in such manner is pattern 3 and its result is not included into the calculation of the overall performance). The compared results seem to indicate that this system is sound, where an average rate of 5.96% for false negative and 1.22% for false positive respectively. Overall, the largest contributor for the false positive is caused from pattern 6, where misclassification of other erroneous patterns into this model. Whereas in terms of false negative, the model that recognizes pattern 5 contributed greatly (see Figure 7 for the graph).

<table>
<thead>
<tr>
<th>Mean start position</th>
<th>Mean ending position</th>
<th>Average sequence length</th>
<th>False positive</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>29th bp</td>
<td>712th bp</td>
<td>684</td>
<td>0.056393</td>
<td>0.243778</td>
</tr>
</tbody>
</table>
6. Conclusion

In this paper, a scheme was advocated for the submission of ESTs data that contain extra information about the signal quality, and the proposed paradigm maybe expect to evolve into a standard format for the ESTs databases in the future. This study addresses a complex and important issue in Bioinformatics that is still have not yet being properly solved and with the growing number of references that indicate various studies are aggressively addressing similar topics [47–50]. Although ESTs are an excellent source of genomic information, there is a consensus that about 3% of the sequence information contains errors in base calling, and this amount is significantly higher than the 0.01% goal specified in the Human Genomic Project. This proposed paradigm seems to has the potential to alleviate this problem, where the error rate is decease to 2%; thus a full modeled system might be able to eradicate the errors in EST data to an acceptable level. The long term goals of this study are to improve the ability and increase the roles of bioinformatics in the development of disease prediction and clinical diagnostic analysis, where the computation results are generated with higher degree of certainty. Furthermore, new erroneous patterns that have not been found in prior literatures have been discovered, which make us believe in that some of the erroneous patterns in ESTs’ trace data might be organism specific, thus such a scheme that is being proposed here is highly suited to solve such problem.

Acknowledgment

This research work was supported by National Science Council of the Republic of China through grant NSC-93-2213-E-182-011, and by Chang Gung Hospital through grant CMRPD33092-8. The basic concepts and ideas of this work have being previously published as a conference paper at International Association for Computer Information Systems Conference (IACIS Pacific 2005), in Taipei, Taiwan.

References

318 Yen-I Chiang and Guan-I Wu


Manuscript Received: Feb. 25, 2008
Accepted: Apr. 25, 2009