Linköping Electronic Articles in
Computer and Information Science
Vol. 6(2001): nr 20

Enumerating Alternative Spliced Transcripts for Understanding Complexity of the Human Genome

Shinichi Morishita, Jun Ogasawara,

Toshihiko Honkura, and Tomoyuki Yamada

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/2001/020/

Published on August 30, 2001 by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 6(2001): nr 20. http://www.ep.liu.se/ea/cis/2001/020/. *August 30, 2001.*

This URL will also contain a link to the author's home page.

Abstract

According to recent several reports, the number of human genes is estimated fewer than 40,000, which, surprisingly, is only double the number (19,000) present in Caenorhabditis elegans. This fact indicates that the correlation between complexity of a species and its gene number appears to be loose, motivating studies on other sources of complexity such as alternatively splicing and cis-regulatory machinery. The first step towards this research direction would be to list candidates of alternatively spliced transcripts and promoters. Because of the release and updates of the human genome, this enumeration could be achieved by aligning millions of expressed sequence tags (ESTs, for short) to a draft human genome and then organizing alignments into related groups, which is however computationally costly. We have developed an efficient algorithm for this purpose, taking only one day to align three millions of ESTs to a newly revised draft genome from scratch. Analysis of 2.2 millions of alignments identifies about 8,000 groups of alternatively spliced transcripts. A typical group contains tens of alignments, demanding a method of selecting representatives. All the results are accessible through the WWW at our Gene Resource Locator web site `http://grl.gi.k.u-tokyo.ac.jp/`.

Contact with authors

Authors' affiliation: University of Tokyo

Contact Person: Shinichi Morishita

E-mail: `moris@k.u-tokyo.ac.jp`

Web page: `http://www.gi.k.u-tokyo.ac.jp/~moris/`

1 Introduction

We first briefly review the transfer of information from DNA to protein in eucaryotic cells, which is illustrated in Figure 1. Most eucaryotic genes have their coding sequences, which are called *exons*, interrupted by non-coding sequences called *introns*. First, the complete length of the gene that includes both its exons and its introns is transcribed into a long RNA sequence called the *primary transcript*. Before this primary transcript leaves the nucleus, a complex of RNA-processing enzymes removes all the introns, yielding a shorter RNA molecule. This RNA-processing step is called *RNA splicing*. After the RNA splicing step has been completed, the RNA molecule moves to the cytoplasm as an *messenger RNA* (*mRNA*, for short) that is further translated into a particular protein.

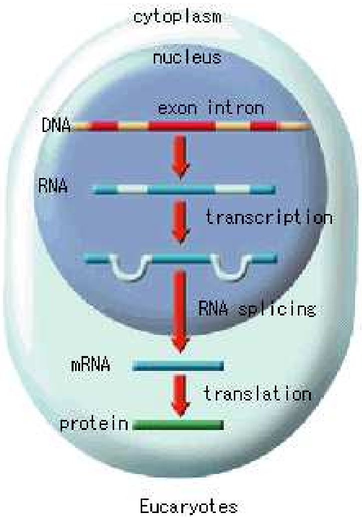

Figure 1: Transfer of information from DNA to protein

The process of reading primary transcripts and then eliminating introns may appear to be wasteful, but the process is essential to synthesize various types of genes, because primary RNA transcripts of a particular gene could be spliced in various ways to produce different mRNAs. This mechanism is called *alternative splicing*, and the objective of this paper is to study an efficient way of listing alternatively spliced transcripts. For this purpose, to date, ample raw data are publicly available; namely, (1) draft sequences of human genome that include more than 90% of the entire nucleotide sequence are accessible through the WWW [3, 4], and (2) about three millions of mRNAs or their fragments called expressed sequence tags (EST, for short) are stored in public databases such as dbEST [1] and UniGene [2]. Thus, aligning millions of ESTs to the draft DNA sequence is useful to identify alternatively spliced transcripts. This computation is however time-consuming, because of the tremendous size of the human genome and the enormous number of ESTs, thereby demanding an efficient solution to this problem.

2 Aligning millions of ESTs to the human genome efficiently

We here overview the crux of our fast alignment algorithm proposed in [8].

2.1 Index generation

The first step in mapping ESTs to the genomic sequence was to build an index where the primary keys (called N-mers) were nucleotide sequences of length N. One N-mer might appear in more than one location in the genome. Therefore, in the index, multiple occurrences of N-mer could be associated with each primary key. The index helped to locate the positions of one N-mer instantly, a crucial step in accelerating the overall performance of mapping millions of ESTs in the genome. Care had to be taken in selecting appropriate values for N. Smaller N values, say 5, aligned an enormous number of positions for each 5-mer, while a larger N value increased the number of N-mers, yielding a huge index. A huge index could be squashed by the hashing technique, although we found this hashing technique introduced additional overhead time to access each N-mer key. Considering that the index is constantly read, we created a simple index instead of an elaborated hash table. We selected 14-mers so that the average number of occurrences of one N-mer was about 12. Since the number of 14-mers is about 256 million, the number of occurrences of 14-mers in the human genome is about 3 billion, and the quotient of the latter figure by the former is about 12.

2.2 EST alignment using the index look-up

The second step was to align one EST to the genome. Our strategy was to approximate the starting and ending positions of each alignment of one EST (*an EST alignment*, for short, in what follows) by mapping 21-mers at each end of the EST. The algorithm scanned the EST from the head until one 21-mer in the EST was fully matched to the genome. Some 21-mers were not immediately evident because of errors in the EST sequences, but in most cases, it was sufficient to search the head 100 bp to find the first match. Similarly, the EST was searched from the tail to find one 21-mer fully matched to the genome. Once the starting and ending positions of the EST alignment were approximated, we extended the starting 21-mer in the EST and the genome simultaneously by looking up the index of 14-mers until the EST to genome match was lost. We then searched the location of the remaining part of the EST by referring to the index to find the starting location of the next exon. At intron boundaries between the two exons, our algorithm moved the windows of two exons so that the exon/intron junctions obeyed the so-called "GT-AG" rule. The entire process was iterated until all the exons were identified. We used Fujitsu PrimePower 1000 with sixteen Fujitsu's SPARC64 GP processors with a CPU clock rate of 450MHz and 32GB of main memory. It took about 10 milli-seconds on average to decide whether or not one EST maps to some locations in the genome, making it possible to process 3 millions of ESTs in less than nine hours with the use of one CPU. Our software has improved the computation time by a couple of orders of magnitude compared to Sim4 [5], which is commonly used in practice.

2.3 Datasets

We used the draft genome assembled from the April 1, 2001 Genbank 11 freeze at the University of California, Santa Cruz, which is accessible through the WWW at `http://genome.ucsc.edu/`. The EST database used was the Unigene Human dataset (Build #136), which contained 2,831,771 ESTs. We selected alignments where: (1) the matching ratio was at least 90%; (2) the coverage ratio was at least 50%; (3) the exon length was at least 16; and (4) the intron length was 400,000 bp maximally. Our software calculated 2.2 millions of alignments of ESTs.

3 Identification of alternative splicing

3.1 EST loci

Millions of EST alignments are redundant, because a number of distinct ESTs might map to the same locus. In this case, alignments sharing common exons in the same strand should be placed in the same group, which is called *EST locus* [7, 6]. In an EST locus, we could frequently observe alternatively spliced transcripts having different combinations of exons.

For instance, Figure 2 illustrates an EST locus. In the figure, each thick line represents the EST alignment in which the narrow yellow boxes are exons and the blue boxes are introns. The bottom thick line, for example, presents the EST alignment of the RefSeq gene numbered NM_007196 (kallikrein 8) of length 1007 bp aligned to q13 in Chr.19 with a matching ratio of 100%. We identify EST loci by iterating the process of grouping two EST alignments in the same strand that share common exons at the same positions until each group converges; Figure 2 shows such an example. The bottom alignment for NM_007196 includes all the exons, and hence it is reasonable to select it as representative.

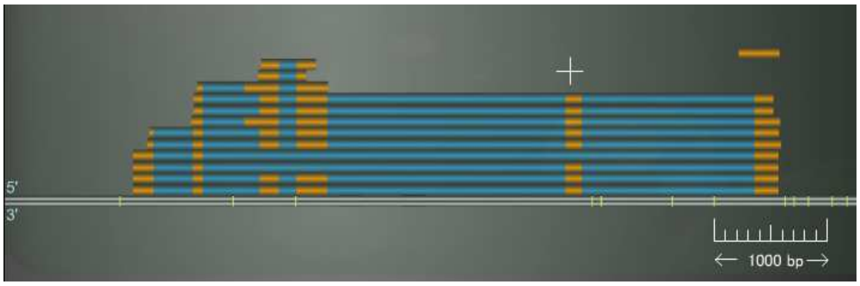

Figure 2: Example of an EST locus

3.2 Alternatively spliced transcripts

Figure 2 reveals some alternatively spliced transcripts. For instance, there are some transcripts that do not use the third, fourth, and fifth exons of NM_007196. In Figure 2, all the EST sequences are aligned from 5' to 3' and displayed from left to right. The intronless alignment at the upper right corner of the figure is a 3' ESTs with poly-adenylation signals and poly-adenine, which confirms the existence of this gene-coding region.

Figure 3 illustrates an EST locus in which no EST can be selected as representative, since no EST covers all the exons and no EST ranges across all the EST coding regions, although the bottom alignment corresponds to the RefSeq sequence NM_017727. This example demonstrates that mapping known genes such as RefSeq sequences is insufficient to fully understand the structures of gene coding regions, thereby demanding many EST alignments as supplementary information.

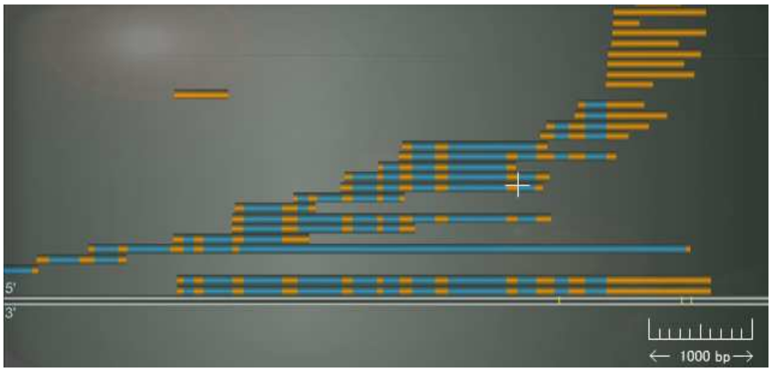

Figure 3: Example of an EST locus

3.3 Subsumption among EST alignments

In Figures 2 and 3, some EST alignments are redundant in the sense that they appear to be subsumed by others. To clarify this point, in Figure 4, to each individual EST alignment, we assign a unique number that increases from lower to higher in the picture.

In Figure 4, EST alignment #13 is intronless, and it is not consistent with other exons at the same position. This EST is suspected be a product owing to DNA contamination. In general, we rule out intronless alignments, because they are not helpful to detect alternatively spliced transcripts.

Next, note that the alignment #1 includes all the exons that appear in other alignments, which might lead us to conclude that all the ESTs function similarly. This conclusion is however misleading, because the alignment #3, for instance, does not use the second and the third exons of #1, coding a different type of protein. The same observation is applied to #4 and #7.

The alignments #2, #5, #6, #8, and #9, on the other hand, use all the exons of #1, and they are shorter than #1. In this case, it is reasonable to treat them as fragments of #1. Also, #10 is a fragment of #7, and #11 and #12 are fragments of #1 and #9.

3.4 Maximal EST alignments

In order to formalize what we have informally called fragmentary relationship between EST alignments, one may consider using the notion of subsequence relationship between original ESTs; for instance, #11 is a fragment

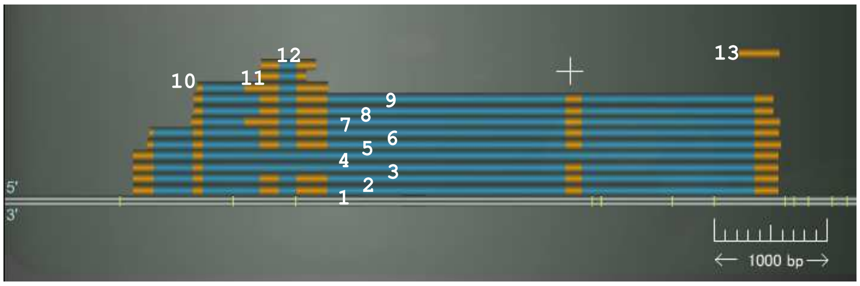

Figure 4: Transcripts with identifiers of numbers

of #1, because the EST sequence for #11 is a sub-sequence of EST for #1. However, this simple definition does not work appropriately, because errors in EST sequences are frequently observed, and the ratio is said to be about a percentage of five. On the other hand, the error ratio of the genomic DNA sequence is smaller than 0.01%. We will therefore take advantage of the accuracy of the genomic sequence, and we will introduce another way of defining subsumption relationship between EST alignments .

The task of aligning an EST sequence $E = e_1, e_2, \ldots, e_N$ ($e_i \in \{a, t, g, c\}$) to the genome $G = g_1, g_2, \ldots, g_M$ ($g_i \in \{a, t, g, c\}$) is twofold;

1. We decompose $[1, n]$ into disjoint sub-ranges expressing exons
$$[1, c_1], [c_1 + 1, c_2], \ldots, [c_{k-2} + 1, c_{k-1}], [c_{k-1} + 1, N]$$
such that $c_i < c_{i+1}$ for $i = 1, \ldots, k - 1$, and their union reproduces the original $[1, N]$.
2. We also generate a mapping $f : [1, N] \rightarrow [1, M]$ such that f preserves the order in $[1, N]$; namely, $f(i) < f(i + 1)$ for $i = 1, \ldots, N - 1$.

Observe that the series of sub-ranges
$$[f(1), f(c_1)], [f(c_1 + 1), f(c_2)], \ldots, [f(c_{k-2} + 1), f(c_{k-1})], [f(c_{k-1} + 1), f(N)],$$
which we will denote $E_1, E_2, \ldots, E_k$, is the alignment of E to the genome G.

Suppose that another EST F is given, and let $F_1, F_2, \ldots, F_l$ denote the alignment of F to the genome. Let $F_1, F_2, \ldots, F_l \leq E_1, E_2, \ldots, E_k$ denote that (or, $F \leq E$ for short) there exists $p \in [1, M]$ satisfying the following conditions:

- The first exon of F is a sub-range of the p-th exon of E; that is, $F_1 \subseteq E_p$.
- Every internal exon of F must coincide with an exon of E, and all the internal exons of F must appear consecutively in F; namely, $F_j = E_{p+j-1}$ for $j = 2, \ldots, l - 1$.
- The last exon of F is a sub-range of the $(p + l - 1)$-th exon of E; $F_l \subseteq E_{p+l-1}$.

In the above definition, if the first and last exons of E are respectively exactly aligned to the first and last exons of F (namely, $F_1 = E_1$ and

$F_l = E_k$), we express this fact by $F_1, F_2, \ldots, F_l = E_1, E_2, \ldots, E_k$ (or, $F = E$ for simplicity).

It is easy to observe that $\leq$ defines a partial order among EST alignments; that is, $F \leq H$ and $H \leq E$ implies $F \leq E$, and $F \leq E$ and $F \geq E$ means $F = E$. Let $F < E$ denote the case when $F \leq E$ but $F \neq E$. In Figure 4, for example, $\#11 < \#8 < \#2 < \#1$ and $\#10 < \#7$.

An EST E is *maximal* if there exists no EST H such that $E < H$. For example, #1, #3, #4, and #7 are maximal in Figure 4. It is reasonable to select maximal EST alignments as the representatives of all the alignments in the same EST locus.

3.5 Statistics

By grouping about 2.2 millions of EST alignments, we have identified 27,369 EST loci having multiple exons. Among those EST loci, 8,091 loci contain more than one maximal EST alignment.

4 Ongoing work

We are now studying a way of investigating maximal EST alignments in more precise. For instance, we consider distinguishing maximal EST alignments having meaningful open reading frames that are expected to code proteins from the others without open reading frames. We are also interested in a method of observing gene expression patterns of maximal EST alignments. Actually, in our WWW server `http://grl.gi.k.u-tokyo.ac.jp/`, we assign BodyMap [9] gene expression patterns in various human tissues to more than ten thousands of EST loci. This association of gene expression patterns is not complete, because we observe expressions using short primers on the last 3'end exons, and hence we may see the mixture of expression patterns of distinct maximal EST alignments. To overcome this problem, we are looking at an efficient way of observing expression patterns of internal exons.

References

[1] http://www.ncbi.nlm.nih.gov/dbest/.

[2] http://www.ncbi.nlm.nih.gov/unigene/.

[3] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–, 2001.

[4] J.C. Venter et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[5] L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. A computer program for aligning a cdna sequence with a genomic sequence. *Genome Research*, 8(9):967–974, 1998.

[6] Toshihiko Honkura, Jun Ogasawara, , Tomoyuki Yamada, and Shinichi Morishita. Grl viewer — a dynamic graphical browser of gene structures and splicing patterns. In *Proceedings of Genome Informatics*. Cold Spring Harbor Laboratory/Wellcome Trust, August 2001.

[7] Shinichi Morishita, Jun Ogasawara, and Toshihiko Honkura. Gene number estimation by aligning millions of expressed sequence tags with the human genome sequence. *Submitted.*

[8] Jun Ogasawara and Shinichi Morishita. Fast algorithm for aligning expressed sequence tags to genomic dna. *Submitted.*

[9] J. Sese, H. Nikaidou, S. Kawamoto, Y. Minesaki, S. Morishita, and K. Okubo. Bodymap incorporated pcr-based expression profiling data and a gene ranking system. *Nucleic Acids Res.*, 29:156–158, 2001.