Report Number: WUCS-2006-25

2006-05-01

Using Expressing Sequence Tags to Improve Gene Structure Annotation

Chaochun Wei and Michael R. Brent

Finding all gene structures is a crucial step in obtaining valuable information from genomic sequences. It is still a challenging problem, especially for vertebrate genomes, such as the human genome. Expressed Sequence Tags (ESTs) provide a tremendous resource for determining intron-exon structures. However, they are short and error prone, which prevents existing methods from exploiting EST information efficiently. This dissertation addresses three aspects of using ESTs for gene structure annotation. The first aspect is using ESTs to improve de novo gene prediction. Probability models are introduced for EST alignments to genomic sequence in exons, introns, interknit regions, splice sites... **Read complete abstract on page 2.**

Using Expressing Sequence Tags to Improve Gene Structure Annotation

Chaochun Wei and Michael R. Brent

Complete Abstract:

Finding all gene structures is a crucial step in obtaining valuable information from genomic sequences. It is still a challenging problem, especially for vertebrate genomes, such as the human genome. Expressed Sequence Tags (ESTs) provide a tremendous resource for determining intron-exon structures. However, they are short and error prone, which prevents existing methods from exploiting EST information efficiently. This dissertation addresses three aspects of using ESTs for gene structure annotation. The first aspect is using ESTs to improve de novo gene prediction. Probability models are introduced for EST alignments to genomic sequence in exons, introns, interknit regions, splice sites and UTRs, representing the EST alignment patterns in these regions. New gene prediction systems were developed by combining the EST alignments with comparative genomics gene prediction systems, such as TWINSCAN and N-SCAN, so that they can predict gene structures more accurately where EST alignments exist without compromising their ability to predict gene structures where no EST exists. The accuracy of TWINSCAN_EST and NSCAN_EST is shown to be substantially better than any existing methods without using full-length cDNA or protein similarity information. The second aspect is using ESTs and de novo gene prediction to guide biology experiments, such as finding full ORF-containing-cDNA clones, which provide the most direct experimental evidence for gene structures. A probability model was introduced to guide experiments by summing over gene structure models consistent with EST alignments. The last aspect is a novel EST-to-genome alignment program called QPAIRAGON to improve the alignment accuracy by using EST sequencing quality values. Gene prediction accuracy can be improved by using this new EST-to-genome alignment program. It can also be used for many other bioinformatics applications, such as SNP finding and alternative splicing site prediction.

2006-25

Using Expressed Sequence Tags To Improve Gene Structure Annotation, Doctoral Dissertation, May 2006

Authors: Chaochun Wei

Corresponding Author: wei@cse.wustl.edu

Abstract: Finding all gene structures is a crucial step in obtaining valuable information from genomic sequences. It is still a challenging problem, especially for vertebrate genomes, such as the human genome. Expressed Sequence Tags (ESTs) provide a tremendous resource for determining intron-exon structures. However, they are short and error prone, which prevents existing methods from exploiting EST information efficiently. This dissertation addresses three aspects of using ESTs for gene structure annotation. The first aspect is using ESTs to improve de novo gene prediction. Probability models are introduced for EST alignments to genomic sequence in exons, introns, interknit regions, splice sites and UTRs, representing the EST alignment patterns in these regions. New gene prediction systems were developed by combining the EST alignments with comparative genomics gene prediction systems, such as TWINSCAN and N-SCAN, so that they can predict gene structures more accurately where EST alignments exist without compromising their ability to predict gene structures where no EST exists. The accuracy of TWINSCAN_EST and NSCAN_EST is shown to be substantially better than any existing methods without using full-length cDNA or protein similarity information.
The second aspect is using ESTs and de novo gene prediction to guide biology experiments, such as finding full ORF-containing-cDNA clones, which provide the most direct experimental evidence for gene structures. A

Type of Report: Other

WASHINGTON UNIVERSITY

THE HENRY EDWIN SEVER GRADUATE SCHOOL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

USING EXPRESSED SEQUENCE TAGS TO IMPROVE GENE STRUCTURE ANNOTATION

by

Chaochun Wei

Prepared under the direction of Professor Michael R. Brent

Dissertation presented to the Henry Edwin Sever Graduate School of

Washington University in partial fulfillment of the

requirements of the degree of

DOCTOR OF SCIENCE

May 2006

Saint Louis, Missouri

WASHINGTON UNIVERSITY

THE HENRY EDWIN SEVER GRADUATE SCHOOL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ABSTRACT

USING EXPRESSED SEQUENCE TAGS TO IMPROVE GENE STRUCTURE ANNOTATION

by

Chaochun Wei

ADVISOR: Professor Michael R. Brent

May 2006

Saint Louis, Missouri

Finding all gene structures is a crucial step in obtaining valuable information from genomic sequences. It is still a challenging problem, especially for vertebrate genomes, such as the human genome. Expressed Sequence Tags (ESTs) provide a tremendous resource for determining intron-exon structures. However, they are short and error prone, which prevents existing methods from exploiting EST information efficiently. This dissertation addresses three aspects of using ESTs for gene structure annotation.

The first aspect is using ESTs to improve de novo gene prediction. Probability models are introduced for EST alignments to genomic sequence in exons, introns, interknit regions, splice sites and UTRs, representing the EST alignment patterns in these regions. New gene prediction systems were developed by combining the EST alignments with comparative genomics gene prediction systems, such as TWINSCAN and N-SCAN, so that they can predict gene structures more accurately where EST alignments exist without compromising their ability to predict gene structures where no EST exists. The accuracy of TWINSCAN_EST and NSCAN_EST is shown to be substantially better than any existing methods without using full-length cDNA or protein similarity information.

The second aspect is using ESTs and de novo gene prediction to guide biology experiments, such as finding full ORF-containing-cDNA clones, which provide the most direct experimental evidence for gene structures. A probability model was introduced to guide experiments by summing over gene structure models consistent with EST alignments.

The last aspect is a novel EST-to-genome alignment program called QPAIRAGON to improve the alignment accuracy by using EST sequencing quality values. Gene prediction accuracy can be improved by using this new EST-to-genome alignment program. It can also be used for many other bioinformatics applications, such as SNP finding and alternative splicing site prediction.

To

My parents XinGuang Zhang, Delin Wei and my wife, Hong Hu

Contents

Tables

Figures

Glossary

AB INITIO METHOD: a collection of gene prediction methods only taking genomic sequences as inputs.

AFFINE: An affine score of length d is $s(d) = -a - (d-1)*b$, while a is called the open penalty and b is the extension or continuous penalty. An affine score is usually used for gap penalty in sequence alignment.

ALIGNMENT: A mapping of two or more sequences.

ALIGNMENT SCORE: A value representing the similarity of aligned bases.

ALIGNMENT SCORE SCHEME: A way to measure the similarity of aligned symbols of the aligned sequences.

AMINO ACID: One of the building blocks of proteins. Each combination of three continuous nucleotides determines one amino acid. There are 20 different types of amino acids corresponding to 64 different combinations of nucleotide sequence of length three.

AMPLICON: The DNA product of a PCR reaction, usually an amplified segment of a gene or DNA.

ANNOTATION: A set of biologically meaningful features of genomic sequences.

BASE: The building block of DNAs and RNAs; also as a Nucleotide.

BASE CALLING: A procedure to obtain DNA sequences from trace files output from sequencing machines.

BACKWARD ALGORITHM: An algorithm that can be used to find the total probability of all the possible paths for a HMM. See Forward Algorithm.

CANONICAL INTRON: An intron that starts with “GT” or “GC” and ends with

“AG”.

CDNA: Complementary DNA. It is a form of DNA prepared in a laboratory with an mRNA as the template by a procedure called reverse transcription (from an mRNA to a DNA).

CDNA CLONE: A cloning vector, which can be a bacterial virus or plasmid, containing a segment of cDNA. A cloning vector can replicate itself and the contained DNA segment in a host bacterium.

CHOROMSOME: A single, long, DNA molecule with its attendant proteins that moves as an independent unit during mitosis and meiosis.

CODON: Three continuous nucleotides encoding an amino acid of a protein.

COMPARATIVE GENOMICS: The study of genetics by comparisons of the genomes of different species. It is still a young field, but has great promise to produce insights into many aspects of the evolution of modern species.

CONSERVATION: A measurement of similarity that remains among different genomes due to evolution under natural selection pressure.

CONSERVATION SEQUENCE: A sequence derived from two genome comparison to present the conservation patterns between these two genomes.

DELETION: A type of genetic mutation by loss of one or more adjacent nucleotides in a DNA sequence.

DNA: Deoxyribonucleic acid; the molecule inside the nucleus of a cell that carries genetic information. DNA is double-stranded.

DYNAMIC PROGRAMMING: A collection of algorithms that solve problems by combining the solutions to sub-problems. Dynamic programming is typically applied to optimization problems.

EST: Expressed Sequence Tag. An EST is a single sequencing read from either end of a cDNA clone; so it is short and error prone.

EXON: A DNA sequence segment in a gene that codes for a transcript product. The usage of the term “Exon” varies in different literatures. In this dissertation it stands for a protein coding region exon. It is called “UTR exon” if it is used to include the non-coding regions of an mRNA in this dissertation.

EXPRESSED: A gene has been activated or “turned on”.

FORWARD ALGORITHM: An algorithm used to find the total probability of all the possible paths for a HMM. It can be used for posterior decoding when used with Backward Algorithm. See Backward Algorithm.

FULL-LENGTH: Containing all the coding region of a gene.

FULL-LENGTH CDNA: A cDNA containing all the transcribed region of a gene.

FULL ORF: Full-length Open Reading Frame; an Open Reading Frame contains all the coding region of a gene. See ORF.

FULL ORF-CONTAINING CDNA: A cDNA containing all the coding region of a gene.

GAP: A symbol to represent an empty base in an alignment.

GENE: A piece of DNA that represents a fundamental physical and functional unit of heredity.

GENE PREDICTOR: A computer program for gene finding.

GENOME: The complete genetic materials for an organism.

GHMM: Generalized Hidden Markov Model; a variation of HMM that each state can emit multiple symbols instead of one symbol a time.

GLOBAL ALIGNMENT: The alignment of sequences from one end to the other. See local alignment.

HMM: Hidden Markov Model; a probability model similar to a finite state machine in which the transitions are probabilistic and each state emits an observation probabilistically too. HMM provides mathematical framework for many computational biology applications, including gene prediction and sequence alignment.

HOMOLOGOUS: (Two sequences are) Similar and from a common ancestor.

INDEL: Insertion and/or deletion.

INSERTION: A type of genetic mutation by introduction of one or more adjacent nucleotides in a DNA sequence.

INTRON: A DNA sequence segment in a gene that is not remained in the mature mRNA of the DNA. Introns separate exons.

KILOBASE: one thousand bases; abbreviated as “Kb”.

LOCAL ALIGNMENT: The best alignment of subsequences of the aligned sequences.

MEGABASE: one million bases; abbreviated as “Mb”.

MGC: Mammalian Gene Collection; a project to find at lease one full ORF-containing cDNA clone for each gene in human, mouse, rat, and cow genomes.

MRNA: Messenger RNA; a single-stranded molecule of ribonucleic acid that is transcribed from a DNA and servers as a template for protein synthesis.

MULTIZ: A multiple genomic sequence alignment program. It can align incompletely sequenced genomes as well as completed genomes. N-SCAN uses MULTIZ alignments to generate alignment sequences.

NUCLEOTIDE: A building block of DNA or RNA, consisting of one nitrogenous base, one phosphate molecule and one sugar molecule.

N-SCAN: A newer version of TWINSCAN with a phylogenetic conservation model that models the conservation patterns among multiple genomes.

OPEN READING FRAME (ORF): A portion of a gene’s sequence that is uninterrupted by a stop codon so that it encode a peptide or protein.

PCR: Polymerase Chain Reaction; a technique for amplifying DNA sequence.

PRIMER: A nucleic acid strand that serves as the starting point for DNA replication. A primer is required because most DNA polymerases can only begin synthesizing a new DNA sequence by adding to an existing strand of nucleotides.

POSTERIOR PROBABILITY: The probability of a hypothesis given the observed data.

PROMOTER: A short region occurring just upstream of the start of transcription.

PSEUDOGENE: A DNA sequence that is similar to that of an active gene. A pseudogene can not be translated into a functional protein.

RANDOM SEQUENCE: A sequence that each base in it is generated randomly according to a certain probability distribution and different bases are generated independently.

REVERSE TRANSCRIPTION: A procedure that transfers genetic information from RNA to DNA. See Transcription.

RNA: Ribonucleic acid, a single-strand molecule similar to DNA, which helps in process of decoding the genetic information carried by DNA.

RNA SPLICING: The procedure to remove introns from pre-mRNA to produce mRNAs for translation into proteins.

SEQUENCE FEATURE: A short piece of sequence with a particular biological meaning.

STOP CODON: One of the three codons: TAA, TAG and TGA.

SUBSTITUTION: The replacing of one nucleotide base with one of the other three bases.

TRANSCRIBE: Convert genetic information from a strand of DNA to an mRNA.

TRANSCRIPTION: The procedure of converting genetic information from a DNA to an mRNA.

TRANSLATE: The procedure of converting genetic information from an mRNA to a protein.

TWINSCAN: A GHMM-based ab initio gene prediction program using similarity information between two genomes as well as the target genomic sequence for gene prediction.

UNTRANSLATED REGION (UTR): A gene region that is transcribed into mRNA but is not used for protein coding.

WAM: Weighted Array Model; a natural generalization of WMM. The difference between WMM and WAM is that in WAM, the probability of generating a nucleotide in position depends on nucleotide(s) in position(s) adjacent to it.

WMM: A probability model for a sequence of a certain length in which a nucleotide can be generated according to a position specific probability and a nucleotide in each position is generated independently.

Acknowledgments

I would like to thank my advisor, Dr. Michael Brent, for his insightful advices throughout these years. I gratefully acknowledge my fellow lab members Mani Arumugam, Jeltje van Baren, Randy Brown, Paul Flicek, Ping Hu, Evan Keibler, Aaron Tenney, Min Wang and Robert Zimmermann for their support and encouragement.

Thanks also to my wife, Hong, for her patience and encouragement throughout the thesis-writing process and for her comments as the first reader of this dissertation.

Chapter 1

Introduction

Sequencing complete genomes is only the beginning of the long march to fully understanding them. One of the crucial steps in this march is to annotate the gene structures of the genomes, from which structural or functional information of genes and proteins can be inferred. With more than a thousand genomes sequenced or being sequenced, what is the current status of gene prediction systems to meet the demand of annotating all these genomes accurately and efficiently? The accuracy of de novo gene prediction systems, which can predict novel genes, has been improved greatly by using multiple-genome comparison, though it is still low for vertebrate genomes like the human genome. Gene prediction systems based on alignments of transcription products, such as mRNAs and proteins, can be very accurate. However, they can only predict genes with evidence. Since it is expensive and time consuming to produce these types of high quality resources, their coverage may be low for many genomes, especially for newly sequenced ones. On the other hand, Expressed Sequence Tags (ESTs), also as transcription products, can be generated quickly and inexpensively. Millions of ESTs have been created for many organisms. However, the challenge is how to effectively combine these fast growing ESTs with the genomic sequence resources for gene structure prediction since ESTs are short and error prone, which prevents many existing methods from using EST information effectively.

The ultimate goal of gene prediction is to construct a system that can predict gene structures from genomic sequences accurately so that no further experimental verification is needed. However, we are still very far away from this goal. Currently, any gene structure predicted computationally needs to be verified experimentally. The most direct experimental evidence for gene structures comes from sequencing full-length cDNA clones and aligning the cDNA sequences back to the genomic sequence (see Section 1.3.1 for definition of a cDNA). Full-length cDNA clones are also very important to other experimental investigations, such as functional analysis of genes and their products. Therefore, many large scale research projects for finding and sequencing full-length cDNAs have been started since the1990s, including the Mammalian Gene Collection (The MGC Project Team). The goal of the MGC project is to find at lease one full ORF-containing cDNA clone for each gene in human, mouse, rat, and cow genomes (Strausberg, Feingold et al. 1999; The MGC Project Team 2004). However, finding full ORF-containing cDNA clones is a time-consuming and expensive procedure. With the current technology, genes that are expressed at low level can not be obtained by finding and sequencing full ORF-containing cDNAs. Therefore, no genome has full ORF-containing cDNAs generated for all of its genes yet.

In 1991, Venter and colleagues developed the Expressed Sequence Tag (EST) (Waterston, Lindblad-Toh et al.) strategy to generate cDNA resources. Expressed Sequence Tags (ESTs) (Adams 1991; Boguski, Lowe et al. 1993) are single sequencing reads from either one or both ends of cDNA clones; they can be generated quickly and inexpensively. The importance of ESTs for full-length cDNA finding and gene structure identification was recognized immediately. A public accessible database of ESTs (dbEST division of GenBank) was set up in 1993 to accept ESTs produced worldwide. Since then, the number of EST sequences has been growing exponentially, from about 22 thousand from 7 organisms in 1993 to more than 32 million from more than 1000 organisms in January 2006.

Since ESTs are generated from cDNA, they contain only coding regions and/or Untranslated Regions (UTRs). Their alignments to genomic sequences indicate the intron-exon structure of the gene they covers. Therefore, they provide tremendous resources for gene identification. However, it is difficult to improve gene prediction accuracy by integrating them into existing non-EST-alignment-based gene prediction systems (Ashburner 2000; Reese, Kulp et al. 2000; Salamov and Solovyev 2000; Yeh, Lim et al. 2001). Because ESTs are short, it means one EST usually does not cover a complete gene. Furthermore, their sequencing error rate is high, which causes an accurate EST-to-genome alignment non-trivial. Also, multiple ESTs aligned to a same genomic region may be in conflict with each other especially when they are from alternative splicing. Furthermore, for most genomes, their existing ESTs only cover a small portion of the total genes they have. Even for extensive EST sequenced genomes, such as human genome, there are still about 15% of genes not covered by any ESTs.

1.1 Goals

This thesis addresses three aspects of how to use ESTs to improve gene annotation. The first is a new approach to integrate EST-to-genome alignments with our de novo gene prediction systems based on multiple genome alignments. The new systems can predict gene structures more accurately where ESTs exist without compromising their ability to predict novel genes. The second is an algorithm to use the gene prediction methods and EST alignment information to guide molecular biology experiment like full ORF-containing cDNA clone selection. The last is a new EST-genomic-sequence alignment algorithm using sequencing quality values of ESTs to improve the accuracy of EST-to-genome alignment.

1.2 Organization

This dissertation contains 6 chapters and two Appendices. For readers with little biology background, appendix A is a good place to start with. Chapter 1 describes the background and the importance of the research presented in this dissertation and chapter 2 covers models and algorithms used in this dissertation. The following chapters are organized according to the three goals described above: chapter 3 covers integrating ESTs to improve gene predictions; chapter 4 describes using EST-genomic-sequence alignment and de novo gene structure prediction to guide biology experiments; and chapter 5 presents a new EST-genomic-sequence alignment algorithm using sequencing quality values to improve the alignment accuracy. From chapter 3 through chapter 5, each chapter is organized to include introduction of the models, the implementation details, the results, and a discussion. While containing one research goal, each of these three chapters may stand alone by itself and connect to the others to show the use of ESTs to improve gene structure annotation. Chapter 6 contains the overall conclusion. It discusses the application of the spliced-alignment in chapter 5, including gene prediction and some potential applications beyond gene prediction.

1.3 Background

This chapter introduces the current-existing gene-prediction methods first. Then, the full ORF-containing cDNA clone selection and EST-to-genome alignment is reviewed briefly.

1.3.1 Gene Prediction Methods

Gene structure prediction has been an active research field for decades. Numerous algorithms have been developed for gene structure prediction based on the data source available at the time. Although its accuracy has been improved greatly in recent years, it is still a challenging problem, especially on vertebrate genomes (Burset and Guigo 1996; Guigó, Agarwal et al. 2000; Stormo 2000; Zhang 2002; Brent and Guigo 2004; Brent 2005). Gene prediction systems can be divided into different generations approximately by the time they were developed. The first generation of gene prediction systems had been developed since early 1980s. They could find approximately the boundary of protein-coding regions and non-protein-coding regions in genomic DNAs, but could not assemble the coding regions into translatable mRNAs. The most frequently used information was from some small regions of particular features or signals, which include splice sites and promoter sites. The compositional difference of coding and non-coding regions, such as codon usage bias, was widely used in the first generation gene prediction programs, too. The most popular first generation gene prediction programs might be TestCode (Fickett 1982), which used base distribution on different codon positions, and GRAIL (Uberbacher and Mural 1991), which used a neural network to combine 7 different types of sequence content information.

The second generation of gene prediction systems could predict the whole gene structure from the start of translation to the stop of translation. They were available in early 1990s. The earliest such programs might be gm (Fields and Soderlund 1990) for *C. elegans* gene prediction and Gelfand method (Gelfand 1990) for mammalian genes. Since then, a number of such programs have appeared, such as GeneID (Guigó, Knudsen et al. 1992), GeneParser (Snyder and Stormo 1993; Snyder and Stormo 1995), FGENSH (Solovyev, Salamov et al. 1994), GRAIL II and GAP (Xu, Einstein et al. 1994). Most of these second generation programs could only predict a single gene on a relative short genomic sequence.

Starting from mid-to-later 1990s, the third generation of gene prediction systems could predict multiple complete or partial gene structures in a long genomic sequence. One of the most successful such programs was GENSCAN (Burge and Karlin 1997), which used a generalized Hidden Markov Model framework for gene prediction. It had dominated in mammalian gene prediction for years until new gene prediction systems based on comparative genomics came on the stage.

The success of comparative-genomic-based gene prediction systems is one of the most important progresses in gene prediction for the past several years. For this type of methods, the growing number of genomes sequenced or being sequenced produces not only a challenge but also a great opportunity. They hold great promise with more genomes sequenced or begin sequence.

Classification of Gene Prediction Methods

Gene prediction approaches could also be classified by the type of the information they used. They can be divided into ab initio, transcript-alignment based and hybrid methods which combines ab initio and transcript-alignment based approaches.

Ab initio Methods

Ab initio methods, such as GENSCAN (Burge 1997; Burge and Karlin 1997) , GeneFinder (Green unpublished) and Fgenesh (Solovyev and Salamov 1997; Salamov and Solovyev 2000), use only the DNA sequence and gene structure models. Information in a DNA sequence includes those characteristic of splice donor and acceptor sites, translation initiation and termination sites, and codon usage within exons.

Transcript-alignment-based Methods

Transcript-alignment-based methods, such as ENSEMBL (Birney, Clamp et al. 2004), use alignments of transcript products, such as proteins, cDNA sequences, mRNA sequences and ESTs, to a genome sequence as the primary basis for gene predictions. One advantage of these methods is that they can predict genes with very high accuracy if those genes have their own transcripts sequenced or they are very similar to some sequenced transcripts. With the enormous increase in the number of known protein coding genes, the accuracy of transcript-alignment-based gene prediction programs has been improved a lot. Since later 1990s, transcript-alignment-based methods have been used routinely for gene prediction (Stormo 2000). However, transcript-alignment-based methods can only predict genes where transcription evidence exits (Birney, Clamp et al. 2004). This is a significant limitation, since sequencing cDNA libraries generally produces only about 50-60% of the genes in a genome (Seki, Satou et al. 2004). Even when genes partially covered by ESTs are included, that number of genes may be only up to 70-85%. cDNA sequences of genes that are expressed at a low level or in a small number of tissues may not be sequenced even after sequencing libraries very deeply (Guigó, Dermitzakis et al. 2003; The MGC Project Team 2004).

Hybrid Methods

A recent trend is toward hybrid methods that combine all the possible information for gene prediction (Allen, Pertea et al. 2004). Hybrid methods that integrate EST, cDNA or protein similarities into ab initio methods have been developed, such as Genie (Kulp, Haussler et al. 1996), Fgenesh++ (Solovyev and Salamov 1997) and Genomescan (Yeh, Lim et al. 2001). Since hybrid methods can use both information from genome sequences and sources other than genome sequences, they have great promise. However, it turned out that it is not easy to improve the prediction accuracy by incorporating EST alignments into the pure ab initio gene prediction methods since ESTs are short and error prone (Reese, Hartzell et al. 2000; Reese, Kulp et al. 2000; Salamov and Solovyev 2000; Parra, Agarwal et al. 2003). Most hybrid methods focus on using proteins, mRNAs and full ORF-containing cDNAs. Some authors found that ESTs are not useful for their gene prediction systems (Salamov and Solovyev 2000; Yeh, Lim et al. 2001). Several authors did present evaluations of the effects of using ESTs alone without using protein sequences (Krogh 2000; Reese, Hartzell et al. 2000; Reese, Kulp et al. 2000; Howe, Chothia et al. 2002; Stanke, Schoffmann et al. 2006). Krogh reported no improvement in predictions for fly genome using HMMGene, a HMM-based ab initio gene predictor and Reese reported an increase in sensitivity accompanied by a smaller decrease in specificity (Reese, Kulp et al. 2000). Better results were reported from a program called GAZE (Howe, Chothia et al. 2002). It obtained both sensitivity and specificity increase on *Caenorhabditis elegans* genome. GAZE takes in features like potential exons generated by an ab initio gene predictor GeneFinder (Green, unpublished) and similarity alignments, as input instead of genomic sequences. It provides frameworks using dynamic programming to combine these features from different sources. More positive results about using ESTs for gene prediction were recently reported from AUGUSTUS+ (Stanke, Schoffmann et al. 2006), a GHMM-based gene prediction program used proteins, ESTs or both as extra sources of hints. AUGUSTUS+ extends AUGUSTUS (Stanke and Waack 2003), a GHMM-based gene prediction program that is very similar to GENSCAN but with more realistic intron length distributions. EST or protein alignments were converted into 6 types of hints, which include translation start site, stop codon, donor splice site, acceptor splice site, coding region and fragment of a coding region; and hints from different sources were assigned with different weights to represent the reliability of the sources. Then the optimal gene structures were computed based on these hints and the target genomic sequence. Only hints on coding regions or boundaries of coding regions can be used in AUGUSTUS+, and hints like ESTs have to be preprocessed to get the coding region information first, which itself may be a tricky procedure. Some usefully information of ESTs may also become lost during this procedure.

More researchers tried to use multiple gene prediction systems to produce different sets of predictions, and then combine the resulting predictions together (Lander, Linton et al. 2001; Rogic, Ouellette et al. 2002; Stein, Bao et al. 2003; Allen, Pertea et al. 2004). The problem with this type of methods is that every source needs to be weighted appropriately and it often can be very complicate, especially when they are adapted for new organisms. Although all these methods could improve the accuracy of gene predictors, it does not mean we should stop improve the underlying gene prediction systems.

The author improved the accuracy of the ab initio gene prediction system TWINSCAN on *C. elegans* by using another worm (*C. briggsae*) as informant, incorporating a more realistic intron length distribution and employing more general splice site models (Wei, Lamesch et al. 2005). The experimental result showed that more than a thousand complete novel genes up to date could be derived by RT-PCR from predicted gene structures. In fact, identifying genes by RT-PCR from predicted gene structures is feasible and has been successfully used for some genomes (Flicek, Keibler et al. 2003; Guigó, Dermitzakis et al. 2003; Wu, Shteynberg et al. 2004; Wei, Lamesch et al. 2005). This puts the de novo methods in a more important position. The limitations of existing methods are also discussed in a recent review by (Brent 2005).

As of April 9, 2005, there were more than 1000 genomes being sequenced, of which 88 were for animals and 33 were for plants. (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html). This fast growing number of genomes being sequenced put efficient and reliable automatic gene prediction systems in high demand. Meanwhile, the sequencing of multiple close-related genomes also produces exciting opportunities for researchers to conduct gene structure prediction based on two or multiple genome alignments. However, the challenge remains concerning how to effectively combine these genomic sequence resources and the fast growing EST sequences for gene structure prediction.

1.3.2 Full ORF-containing cDNA Clone Selection

All gene structures predicted computationally need to be verified experimentally. A full-length cDNA clone provides the best experimental evidence of a gene structure. It is also a particularly powerful material for functional study of a gene and its corresponding protein. In the MGC project, candidate cDNA clones were chosen based on their 3' and 5' ESTs, then they were fully sequenced to high accuracy, which is a time consuming and expensive procedure compared to EST sequencing. Because of the polyA tail next to the transcription end, reverse transcription mostly starts from the 3' end. It may stop before it goes through the start of translation. 5' ESTs can be sequenced to evaluate if the cDNA contains the start of translation. In the MGC project, each 5' EST was aligned to RefSeq sequences (Tatusova, Karsch-Mizrachi et al. 1999; Pruitt, Katz et al. 2000; Pruitt, Tatusova et al. 2005), protein sequences and gene predictions by GENOMESCAN (Yeh, Lim et al. 2001) to evaluate if it contains the start of translation. If one of the alignments verified that this 5’ EST contained a start of translation, the cDNA clone of this EST would be selected to be fully sequenced to high accuracy. However, in this target selection method used in MGC project, the property that ESTs contain only coding regions or UTRs was ignored in the gene prediction step. On the other hand, the algorithm developed here can integrate EST alignments into a gene prediction algorithm instead of using them only in post-gene-prediction step. In this way, the sensitivity and specificity of target clone selection could be improved, and this could eventually reduce more cost and time for full ORF-containing-cDNA finding. The algorithm developed here can also be extended to other biology experiments which are expensive and/or time consuming or both. Many unsuccessful experiments may be avoided by using highly accurate gene prediction models to guide the experiments.

1.3.3 EST-to-genome Alignment

The accuracy of EST-genome alignment is critical to many bioinformatics applications, including the above two applications, gene prediction and full ORF-containing-cDNA finding. The high error rate of EST sequences makes the EST-genome alignment algorithm non-trivial. Any increase in the alignment accuracy will improve our understanding of genomic biology procedures such as splicing and consequent wet lab experiment results as well. A sequencing quality value is value assigned by the base-calling program to each base of a sequence to represent the error probability of the base. The quality value of a base shows the reliability of this base. However, the quality values of EST sequence were omitted for all existing EST-to-genome alignment programs. The algorithm developed here is a new method that improves the alignment quality by integrating the quality values of EST sequences with EST-to-genome alignments. Eventually, it will improve the result of gene prediction as well as other computational biology applications.

High quality EST to genome alignment is critical for many computational biology applications. The new method present here can combine the sequencing quality values into spliced alignment to improve the alignment quality

Chapter 2

Models and Algorithms

Chapter 2 introduces the models and algorithms used frequently in this dissertation. The chapter starts with a brief introduction of Markov model, Weighted Matrix Model (WMM) and Weighted Array Model (WAM). Then Hidden Markov Models (HMMs), the probability models of gene structure, and algorithms for them are reviewed in Section 2.2. Section 2.3 covers generalized Hidden Markov Models, and their differences with standard HMMs. Section 2.4 and 2.5 introduce two de novo comparative-genomics-based gene prediction systems TWINSCAN and N-SCAN respectively. A simple pair HMM model for cDNA-genomic sequence alignment will be introduced in section 2.6. Section 2.7 reviews Graphical Model briefly. For readers familiar with models and algorithms in gene prediction, this chapter can be skipped. Otherwise, this can be a tutorial chapter or reference for models and algorithms frequently used in gene prediction.

2.1 Markov Model, WMM and WAM

In computational biology, Markov Model, WMM and WAM have been widely used to modeling genomics sequences and sequence signals, such as the splice donor sites and splice acceptor sites.

2.1.1 Markov Model

Consider a system with N distinct states, $\{1, 2, \dots, N\}$, and it can be described at any time by being in one of the states. At evenly spaced, discrete times, the system changes from state to state with a certain probability associated with the states. If we denote the state at time t as q_t, a full probabilistic description of this system would be $\Pr(q_t \mid q_{t-1}, q_{t-2}, ..., q_1)$, which means the probability of the system staying in the current state depends on all the predecessor states. Such a system can be simplified if the probability of the system being in the current states only depends on a limited number of predecessor states. If the probability of being in the current state only depends on a limited number (*k*) of previous states, the system is called a Markov Chain, and it can be represented by a Markov Model

$$\Pr(q_t \mid q_{t-1}, q_{t-2}, ..., q_1) = \Pr(q_t \mid q_{t-1}, ..., q_{t-k}) \qquad (2.1).$$

The number *k* is called the order of the Markov Model. For example, in a first-order Markov Model, $\Pr(q_t \mid q_{t-1}, q_{t-2}, ..., q_1) = \Pr(q_t \mid q_{t-1})$, the probability of being in the current state only depends on its previous state. If this property is independent of time *t*, i.e.

$$\Pr(q_t = j \mid q_{t-1} = i) = a_{ij} \qquad (2.2),$$

where a_{ij} is a constant for all *t* with fixed *i* and *j*, then this system can be called a homogenous Markov Chain. It can be fully described by state transition probabilities a_{ij}, $1 \le i, j \le N$ and initial probabilities $e_i = \Pr(q_0 = i)$, $1 \le i \le N$. The state transition probabilities a_{ij}, $1 \le i, j \le N$ have the properties $a_{ij} \ge 0$, $\forall 1 \le i, j \le N$ and $\sum_{j=1}^{N} a_{ij} = 1$, $\forall 1 \le i \le N$. Similarly, the initiation probabilities also have the property $e_i \ge 0$, $\forall 1 \le i \le N$ and $\sum_{i=1}^{N} e_i = 1$. In a Markov Chain, if the value of equation (2.1) depends on the position or time *t*, this system may be represented by an inhomogeneous Markov Model.

Weighted Matrix Model (WMM) was introduced by Staden (Staden 1984) to represent sequence signals. It is also called position specific score matrix (PSSM), which was introduced by Stormo (Stormo, Schneider et al. 1982). As its name indicates, under a PSSM (or a WMM), a state is assigned for each position and it can emit symbols according to a state specific probability distribution. When a PSSM or WMM is applied to a sequence of length l, a nucleotide in the sequence can be generated according to the probability defined by the state for each position, and nucleotides in different positions are generated independently. The probability of generating a sequence $x = x_1,...,x_l$ under a WMM model λ with length l is $\Pr(x \mid \lambda) = \prod_{i=1}^{l} \Pr_{(i)}(x_i)$, where $\Pr_{(i)}(x_j)$ is the probability of generating nucleotide x_j at position *i*. Generally $\Pr_{(i)}(x_j)$ can be estimated from $f_{(i)}(x_j)$, the frequency of base x_j at position *i,* by

$$\Pr_{(i)}(x_j) = \frac{f_{(i)}(x_j)}{\sum_{x_j \in A} f_{(i)}(x_j)},$$

where A is the alphabet set containing possible nucleotides of a genomic sequence.

Weighted Array Model (WAM) is a natural generalization of WMM. The difference between WMM and WAM is that in WAM, the probability of generating a nucleotide in position *i* depends on nucleotide(s) in position(s) adjacent to it, while in a WMM, nucleotides in different positions are generated independently. A WMM can be considered as a zero order inhomogeneous Markov Model while WAM is a non-zero order inhomogeneous Markov Model. For example, if λ is a 1st-order WAM , the probability that a sequence is generated under this model will be $\Pr(x \mid \lambda) = \Pr_{(1)}(x_1) \pi_{i=2}^{l} \Pr_{(i-1,i)}(x_i \mid x_{i-1})$, where $\Pr_{(i-1,i)}(x_k \mid x_j)$ is the probability of generating nucleotide x_k at position i given x_j at position $i-1$. In general, it can be estimated by $\Pr_{(i-1,i)}(x_k \mid x_j) = f_{(i-1,i)}(x_j, x_k) / f_i(x_k)$, where $f_{(i-1,i)}(x_j, x_k)$ stands the frequency of $x_j x_k$ on position $(i-1, i)$ and $f_i(x_k)$ is the frequency of x_k at position *i*. $\Pr_{(1)}(x_1)$ at the first position is generated as in WMM.

WMM and WAM have been used to represent sequence signals, such as donor and acceptor splice site. In practice, they are used frequently in the form of a ratio of two probabilities from two models. For example, when a WAM is used for an acceptor splice site, one model is derived from true signals $\Pr^{+}_{WAM}(x)$ and another model $\Pr^{-}_{WAM}(x)$ is derived from those "pseudo-signals", which are not acceptor spice sites but with similar patterns. The ratio of probabilities of two models can be used to discriminate the preference of a sequence staying in a real site or a pseudo-site.

In Chapter 3 of this dissertation, Markov Models are used to represent EST alignment patterns for UTRs, coding regions, intron regions and intergenic regions. WMMs and WAMs are used to represent EST alignment patterns of splicing donor site and acceptor site respectively.

2.2 Hidden Markov Models (HMMs)

The theory of Hidden Markov Models was introduced and studied in the late 1960s and early 1970s. It was implemented for speech processing applications in the 1970s. One of the best tutorial papers about HMM is by (Rabiner 1989). HMMs were introduced to biological applications in the later 1980s (Churchill 1989) for DNA sequence composition analysis and in the early 1990s for many biology applications, such as gene prediction (Krogh, Brown et al. 1994; Krogh, Mian et al. 1994) and protein modeling (Krogh, Brown et al. 1994). A standard Hidden Markov Model (HMM) includes an alphabet set, a set of states, a transition probability matrix describing how to move among the states, state specific probabilities with which a symbol is generated by a state at each time, and a set of initial probabilities for the states. Let λ stand for this HMM, and A be the alphabet set, *x* be the observation sequence, and π be a sequence of states, which can also be called a path. Let π_i be the i^{th} state in the path π and $a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$ the transition probability moving from state *k* to state *l* at position *i-1* to *i*. State *k* emits a symbol *b* from set A with a probability $e_k(b) = P(x_i = b \mid \pi_i = k)$. Therefore, the joint probability of an observed sequence *x* and a state sequence π becomes $P(x, \pi \mid \lambda) = e_{\pi_0} \prod_{i=1}^{L} (e_{\pi_i}(x_i) a_{\pi_{i-1}\pi_i})$, where *L* is the length of the observation sequence *x*, and e_{π_0} is the initial probability of the state at position 1. Given an observation *x* and a HMM λ , the most likely path can be derived by maximizing the probability $P(x, \pi \mid \lambda)$ using Viterbi algorithm (Viterbi 1967), $\pi^* = \arg\max_{\pi} P(x, \pi \mid \lambda)$.

2.2.1 Viterbi Algorithm

Viterbi Algorithm is a variation of dynamic programming. Let *N* be the number of states of an HMM λ, $v_k(i)$ be the probability of the most probable path ending at state *k* with observation x_i at position *i*. Suppose $v_k(i)$ is known for all the states $1 \le k \le N$, then $v_j(i+1) = e_j(x_{i+1}) \max_k (v_k(i) * a_{kj})$. By saving the pointers pointing backwards, the real state sequences can be found by backtracking from the last position *L*. Here is the Viterbi algorithm,

Initialization (*i*=0): $v_k(0) = e_k$ for all k, where e_k is the initial probability of state k, and $\sum_k e_k = 1$.

Recursion (*i*=1,…,L): $v_j(i) = e_j(x_i) \max_k (v_k(i-1) * a_{kj})$;

$$ptr_j(i) = \arg\max_k (v_k(i-1) * a_{kj}) .$$

Termination: $P(x,\pi \mid \lambda) = \max_k (v_k(L))$;

$\pi_L^* = \arg\max_k (v_k(L))$.

Trace back(i=L,…,1): $\pi_{i-1}^* = ptr_i(\pi_i^*))$.

Viterbi algorithm should always be done in a log space in order to avoid underflow problem when calculating $v_j(i)$. The multiply operations above become additions in a log space.

2.2.2 Forward Algorithm, Backward Algorithm and Posterior Decoding

In HMM, the forward algorithm and backward algorithm are used to find the total probability of all the possible paths $P(x \mid \lambda) = \sum_\pi P(x,\pi \mid \lambda)$, and the posterior probability $P(\pi_i = k \mid x, \lambda) = \frac{P(\pi_i = k, x \mid \lambda)}{P(x \mid \lambda)}$ at each position of an observed sequence. Decoding using posterior probabilities has some advantages such as that all the paths are considered. This is important especially when multiple sub-optimal paths have very similar probabilities with the optimal path.

Let $f_j(i)$ be the probability of observing x_1 to x_i with state j at position i, i.e., $f_j(i) = \Pr(x_1,...,x_i,\pi_i = j \mid \lambda)$, then $P(x \mid \lambda) = \Pr(x_1,...,x_L \mid \lambda) = \sum_k \Pr(x_1,...,x_L,\pi_L = k \mid \lambda) = \sum_k f_k(L)$. The forward algorithm to compute $P(x \mid \lambda)$ is

Initialization (i=0): $f_k(0) = e_k$ for all k, where e_k is the initial probability of state k, and $\sum_k e_k = 1$.

Recursion (i=1,…,L): $f_j(i) = e_j(x_i)\sum_k (f_k(i-1) * a_{kj})$.

Termination: $P(x \mid \lambda) = \sum_k (f_k(L))$.

Backward algorithm is analogous to the forward algorithm, but starting from the end of the sequence. Let $b_j(i)$ be the probability of observing x_{i+1} to x_L with state j at position i, i.e., $b_j(i) = \Pr(x_L, ..., x_{i+1}, \pi_i = j \mid \lambda)$. The backward algorithm $P(x \mid \lambda)$ is

Initialization (i=L): $b_k(L) = a_{k0}$ for all k, where a_{k0} is probability of state k going to the end state.

Recursion (i=L-1, …, 1): $b_j(i) = \sum_k e_k(x_{i+1})(b_k(i+1) * a_{jk})$.

Termination: $P(x \mid \lambda) = \sum_k (a_{0k} e_k(x_1) b_k(1))$.

Forward and backward algorithms also need to be done in log space in order to avoid underflow errors. The method for summation of probabilities under log space will be described in section 2.2.3.

All the three algorithms have the same complexity $O(N^2 L)$, where N is the number of states in the HMM, and L is the length of the observation.

The posterior probability $P(\pi_i = k \mid x, \lambda)$ is the probability of state k at position i given the observation sequence x and model λ. Based on results of the forward and backward algorithms, $P(\pi_i = k \mid x, \lambda)$ can be calculated from $f_k(i)$ and $b_k(i)$ with k from 1 to N.

$$P(\pi_i = k \mid x, \lambda) = \frac{P(\pi_i = k, x \mid \lambda)}{P(x \mid \lambda)} = \frac{P(\pi_i = k, x_1, ..., x_L \mid \lambda)}{P(x \mid \lambda)}$$

$$= \frac{P(\pi_i = k, x_1, ..., x_i \mid \lambda) P(\pi_i = k, x_{i+1}, ..., x_L \mid \lambda)}{P(x \mid \lambda)} = \frac{f_k(i) b_k(i)}{P(x \mid \lambda)}$$

$$= \frac{f_k(i) b_k(i)}{\sum_k f_k(i) b_k(i)}$$

HMMs are widely used in bioinformatics since they are analogy to many biology problems intuitively. For example, *x* can be a genomic sequence and the π can be a gene structure, a_{ij} is the probability of a transition between two states, which stand for two different regions in a gene structure; and $e_k(b)$ is the probability that a state *k* generates a symbol *b*, which can be “A”, “C”, “G” or “T” for a genomic sequence. State *k* can be a UTR, exon or intron. The gene structure with optimal probability then can be obtained by optimizing $P(x, \pi \mid \lambda)$ with the Viterbi algorithm.

Posterior probability decoding can be extended further by dividing states into different categories. For instance, in gene prediction, states can be divided into two categories: coding exon or not. Let function g(*k*) = 1 if *k* is the coding exon state, and 0 otherwise. Then $G(i \mid x, \lambda) = \sum_k g(k) P(\pi_i = k \mid x, \lambda)$ is the probability of coding exon state at position *i* given the sequence *x* and model λ. Similar idea can be applied to EST-to-genomic alignment to evaluate the quality of alignments discussed in Chapter 5.

Forward algorithm can be modified to compute only the total probability of a certain type of paths. In Chapter 4 of this dissertation, the forward algorithm is modified to include only paths that are consistent with an EST-genome alignment. Those paths are further divided into two types, and probabilities are then calculated for these two types of paths respectively.

2.2.3 Probability Addition in Log Probability Space

As in Viterbi algorithm, both forward and backward algorithm should always be done in a log space to avoid underflow problem when calculating $f_j(i)$ and $b_j(i)$ for any $1 \le i \le L$ and $1 \le j \le N$. However, the difference is that it is impossible to calculate a logarithm of summation of probabilities from the log value of the probabilities without using the exponentiation and log function. Although both exponentiation and log function are expensive in computation, these can be avoided in practice. Let p_1 and p_2 be two probabilities and $r = p_1 + p_2$. We want to compute $\tilde{r}$ from $\tilde{p}_1$ and $\tilde{p}_2$ where $\tilde{r} = \log_2 r$, $\tilde{p}_1 = \log_2 p_1$ and $\tilde{p}_2 = \log_2 p_2$. We have

$\tilde{r} = \log_2 r = \log_2(p_1 + p_2) = \log_2(2^{\tilde{p}_1} + 2^{\tilde{p}_2}) = \tilde{p}_1 + \log_2(1 + 2^{(\tilde{p}_2 - \tilde{p}_1)})$. So, $\tilde{r}$ can be calculated by adding $\tilde{p}_1$ and $\log_2(1+2^d)$ together, which can be pre-computed and stored in a lookup table with $d = \tilde{p}_2 - \tilde{p}_1$. With a certain level of accuracy, the log value of summation of probabilities from the log values for the probabilities can be finished by a comparison, a subtraction, a table lookup and an addition.

2.3 Generalized Hidden Markov Models (GHMMs)

Unlike a standard HMM, in which each state can emit one symbol each time, each state in a generalized Hidden Markov Model (GHMM) (Rabiner 1989; Stormo and Haussler 1994; Kulp, Haussler et al. 1996; Burge and Karlin 1997) generates multiple symbols according to its state-specific sequence model and its state-specific length distribution model. The first GHMM for gene prediction was introduced in mid-1990s (Stormo and Haussler 1994) and the term Generalized Hidden Markov Model was first used by (Kulp, Haussler et al. 1996). The state specific sequence model is analogy to the emission probability in a standard HMM. A length distribution model replaces the self-loop transition probability of each state in a standard HMM. It controls how long a sequence will stay in this state. The sequence model and length distribution can be very flexible, which makes GHMM much more powerful to model complex systems. If all the state-specific length distribution models are geometric, then the Viterbi algorithm can decode the GHMM very efficiency (linear time in worst case). GENSCAN (Burge and Karlin 1997) is a GHMM-based gene prediction program. However, the exon states in GENSCAN use explicit duration length, which causes the computing time to be proportional to be the cube of the input sequence length theoretically. In practice, the number of exon only grows linearly after the initial several kilo bases because constrains like splicing site signals, translation initiation(ATG), translation termination and in frame stop codons reduce the number of potential exons. With the maxim length limit for the exon states, the running time for GENSCAN becomes $O(N^2D^2L)$, where N is the number of states, D is the limit of exon length and L is length of the input sequence.

2.4 TWINSCAN

TWINSCAN (Korf, Flicek et al. 2001) is a GHMM-based gene prediction program Figure 2-1. It uses the similarity information between two genomes as well as the intrinsic information of the sequence for gene prediction. Instead of aligning genomic sequence to known transcription sequences (cDNAs, mRNAs, proteins or ESTs), genomic sequences are compared to genomic sequence from a different organism. It is a de novo method since it uses only genomic sequences. The sequence models of TWINSCAN were based on GENSCAN models. Conservation sequences are derived from two genome comparison to present the conservation patterns between these two genomes. For a given genomic sequence, TWINSCAN needs a database of sequences from another genome, which is called the informant database, to generate its conservation sequence first. Under the TWINSCAN model, for a given parse, the probability of the genomic sequence and the probability of the conservation sequence generated by a separate conservation model are treated as independent to each other and are combined by multiplying them together. The optimal joint probability is computed by the Viterbi algorithm.

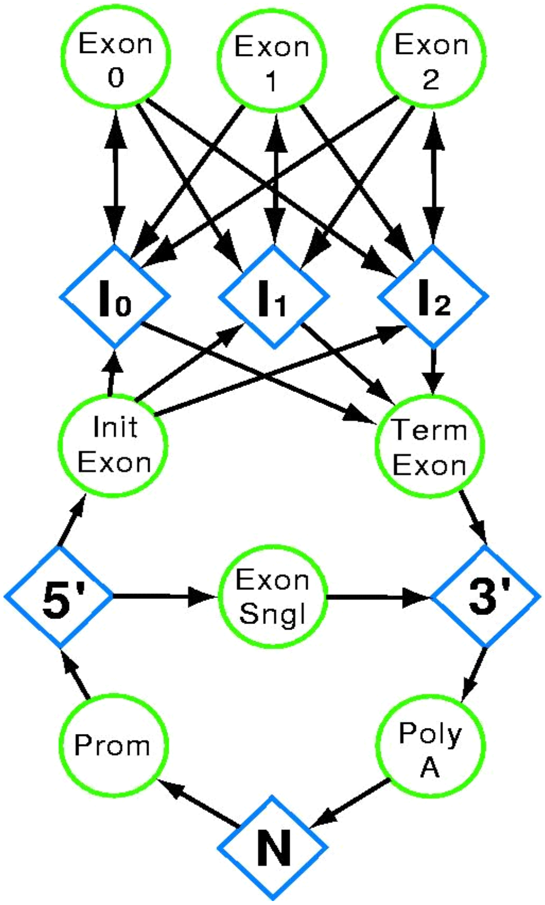


Figure 2-1. TWINSCAN model diagram for the forward strand.

Arrows stand for non-zero probability transitions between states. Exon 0, Exon 1 and Exon 2 stand for internal exons with different reading frame; I0, I1 and I2 represents intron regions with different frame; 5' is for 5' UTR and 3' is for 3' UTR; Prom represents the promoter region; PolyA represents the polyadenylation signal and N represent the intergenic region. The states show here are only for the forward strand, and an analogous model is used fore the backward strand. This figure is from Korf, Flicek et al. 2001.

For the conservation sequences in a TWINSCAN model, separate homogenous 5th-order Markov chains are used for coding regions, UTR, and intron regions. Models of conservation sequence at splice donor sites and acceptor sites are based on two separate 2nd-order Weight Array Matrix models respectively. TWINSCAN's performance improvement is mostly attributed to the homology information from genome comparison. It is one of the most successful gene prediction methods using information derived from genome comparisons (Korf, Flicek et al. 2001; Waterston, Lindblad-Toh et al. 2002; Flicek, Keibler et al. 2003; Guigo, Dermitzakis et al. 2003; Stein, Bao et al. 2003; Wu, Shteynberg et al. 2004; Wei, Lamesch et al. 2005). The original TWINSCAN does not take any transcript similarity evidence. The ideal of combining EST alignments with TWINSCAN is to take advantage from both comparative genomics and EST similarity information.

2.5 N-SCAN

N-SCAN (Gross and Brent 2005) is a newer version of TWINSCAN with a phylogenetic conservation model that models the conservation patterns among multiple genomes, which includes the dependencies between the aligned sequences, context-dependent substitution rates and insertions and deletions in the sequences. The concept of "alignment sequence" was introduced for N-SCAN similar to the "conservation sequence" for TWINSCAN. One alignment sequence is constructed from alignments to each informant genome. N-SCAN allows introns in 5' UTRs. Conserved non-coding regions are added inside the original intergenic region. Several whole-genome-scale gene prediction results showed that N-SCAN could predict gene structure more accurately than any existing ab initio methods, which only use genomic sequences as input. For human genome, the gene sensitivity is about 35% at gene level, compared to 10% and 25% for GENSCAN, TWINSCAN respectively. Again, as for TWINSCAN, there is no transcript similarity information exploited for N-SCAN.

2.6 Pair Hidden Markov Model

The major difference between a standard HMM and a pair-HMM is that each state in a pair HMM generates a pair of symbols instead of a single symbol in the standard HMM. A pair HMM can be used for sequence alignment. Figure 2-2 presents a pair HMM for global pair-wise sequence alignment. Each state (big circles) will emit a pair of symbols. For example, M stands for match or mismatch, and (x_i, y_j) will be emitted from M state every time. X state stands for an insertion in x sequence, so (x_i ,-) will be emitted from state X. Y state is for an insertion in y sequence, and (-, y_j) will be emitted from a Y state. "-" stands for a blank symbol. The best alignment of the sequence x and y can be calculated by Viterbi algorithm.

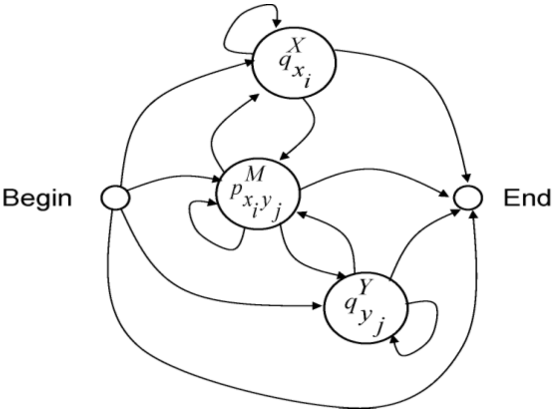


Figure 2-2. Pair HMM for sequence alignment.
M state is for match and mismatch, X state is for the gap in sequence y, and Y state is for the gap in sequence x.

2.7 Graphical Model

A Graphical Model is a combination of graph theory and probability theory. It provides a natural tool to deal with uncertainty and complexity. Its graph parts provide intuitive representation of modularity of a system, in which a complex system is composed of some smaller and simpler systems; and its probability parts provide ways to combine its parts. It has been applied broadly in fields like statistics, system engineering, information theory, pattern recognition and machine learning due to its intuitive interface and ability to represent complex systems.

A graph is a pair *G= (V, E)*, where *V* stands for a set of vertices $\{v_1,...,v_n\}$, and *E* stands for the set of edges, which is a subset of $V \times V$. Let *v* and *w* be two vertices in *V* connected by an edge. If both (*v*, *w*) and (*w*, *v*) are in *E*, then the edge is undirected and it can be represented by a line. If (*v*, *w*) is in *E* while (*w*, *v*) is not in *E*, then the edge is directed and it can be represented by an arrow pointed from v to w. A directed graph contains only directed edges. A directed edge from node *v* to *w* indicates *v* "causes" *w*, and a conditional probability of *w* given *v* can be assigned to this edge. The overall joint probability of a directed graph can be computed as a product of a series of conditional probabilities. Direct graphical model is also called Bayesian Network in many literatures. An undirected graph contains undirected edges only. An undirected edge stands for association of two nodes. A graph containing both direct and undirected edges is called a chain graph if all nodes in this graph can be divided into numbered clusters such that all edges inside the same cluster are undirected and all edges between the clusters are directed, pointing from the set with lower number to the one with higher number. An undirected graph is a special case of a chain graph and so is a directed acyclic graph. By Graphical Chain Model theory (Lauritzen 1996), the overall joint probability of a chain graph can be computed as a production of a series of conditional probabilities of node clusters.

Figure 2-3 shows a simple example of a graphical chain model representing the error sources in EST sequencing. There are 4 nodes in Figure 2-3, "RG", "EG", "EC" and "qual". "RG" stands for the reference genome, from which a genomic sequence is derived and "EG" the EST genome, from which an EST sequence is derived. "EC" is for EST base calls, which are the observed EST bases, and "qual" is for the quality values of the base calls. When both "RG" and "EG" are from the same organism, they are not independent to each other. In other words, they can decide each other.

Therefore, their relationship is represented by a line instead of an arrow. The source EST genome can decide the EST base calls if there is no sequencing error. As a result, the relation between “EG” and “EC” is causal, which is represented by an arrow from “EG” to “EC”. The overall joint probability of this chain graph can be expressed as

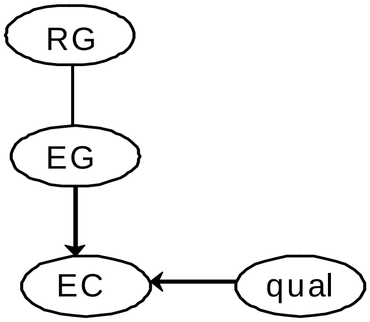


Figure 2-3. A Graphical Model example.

“RG” stands for the reference genome, “EG” for the EST genome, “EC” for EST base calls, and “qual” for quality values of the base calls. When both "RG" and "EG" are from the same organism, they are not independent to each other.

$$\begin{aligned}&\Pr(EC, RG, EG, q) = \Pr(EC \mid RG, EG, q)\Pr(RG, EG \mid q)\Pr(q)\\ &= \Pr(EC \mid EG, q)\Pr(RG, EG)\Pr(q)\end{aligned}$$

An important probability is the probability of observing RG and EC given quality value q. It can be calculated as

$$\Pr(EC, RG \mid q) = \sum_{EG}\Pr(EC, EG, RG \mid q) = \sum_{EG}(\Pr(EC \mid EG, q)\Pr(RG, EG)) .$$

This graphical model is used to investigate the theoretical behavior of the spliced sequence alignment algorithm developed in Chapter 5.

Chapter 3

Using EST Alignments to Improve Gene Prediction

This Chapter describes a new approach to integrate EST alignment information with GHMM-based, de novo gene predictors. When used with comparative genomics based gene-prediction programs, such as TWINSCAN and N-SCAN, it can be automatically retrained to work well on both *C. elegans* and human. Furthermore, the accuracy of TWINSCAN and N-SCAN on genes without EST evidence is not compromised. On the contrary, genes without ESTs are predicted more accurately as a result of the constraints imposed by ESTs aligned to neighboring genes. The concept of ESTseq, which represents the patterns of EST sequence alignment, is introduced in Section 3.1. Construction of an ESTseq from a set of EST alignments is also described in the same section. Section 3.2 covers the models used for ESTseqs and how they are incorporated into TWINSCAN and N-SCAN. Section 3.3 contains parameter estimation for ESTseq models. Section 3.4 reviews measures of the gene-structure-prediction accuracy. Section 3.5 shows the accuracy improvement achieved by combining EST alignments with GHMM-based de novo gene-prediction programs, such as TWINSCAN and N-SCAN. The experimental validation of N-SCAN_EST's novel predictions on the human genome (build NCBI35) is discussed in Section 3.6. Sections from 3.7 to 3.11 investigate the behaviors of the above-mentioned gene-prediction programs using EST alignments. Section 3.7 covers the effect of EST coverage on TWINSCAN_EST prediction accuracy using *C. elegans* as an example. Section 3.8 addresses the trainability of TWINSCAN_EST by comparing its performance when EST alignment parameters are trained from different organism. Section 3.9 covers gene prediction using cross-species EST alignments. Section 3.10 discusses the effect of different alignment programs on the prediction accuracy. Section 3.11 covers the improvement for 5'UTR prediction of N-SCAN_EST by using the EST alignments. This Chapter is ended with a discussion about the importance of the new programs in a more general gene annotation pipeline. The major content of this Chapter has been covered in (Wei and Brent Submitted).

3.1 ESTseq Construction

In order to integrate EST-alignment information into a GHMM-based gene-prediction system, a concept of ESTseq was invented by the author. ESTseq is a sequence with one letter for each base of the input genome which represents much of the useful information in the EST alignments. The method introduced here for exploiting EST alignment is very similar to the "conservation sequence" approach TWINSCAN uses to exploit genomic alignments (Korf, Flicek et al. 2001; Flicek, Keibler et al. 2003). In particular, all available EST sequences are aligned to the genome and alignments that fail to meet certain criteria are filtered out. Each nucleotide of the genome sequence is then assigned one of the three symbols (Figure 3-1): I if it is in an intron of all overlapping EST alignments, E if it is in the exon (aligned region) of all overlapping EST alignments, and N if there is a disagreement among overlapping EST alignments or there is no EST alignment. This representation of EST alignments is called ESTseq by analogy to the conservation sequence or conseq that TWINSCAN uses for genomic alignments. Different strategies have been tried by the author to generate EST-alignment sequences, such as number of symbols used to represent the alignment patterns. For example, an additional symbol U, standing for "unknown", can be used to separate the conflict regions and unaligned regions (N). More symbols can be introduced further to discriminate regions on different strands. The results showed that the simple 3-symbol EST-alignment sequence model described above performs quite well.

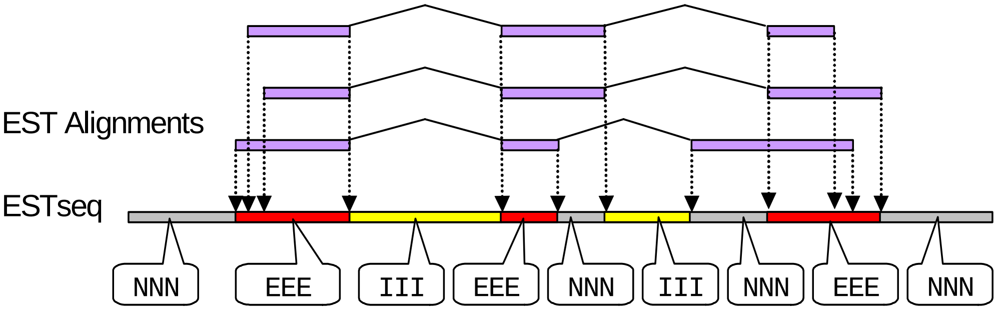


Figure 3-1. Construction of ESTseq from EST alignments.

Each row of top three bars represents the aligned blocks of one EST, while the thin lines connecting the bars represent implied introns. The ESTseq representation contains an "E" for each base that is indicated as exonic (red), an "I" for each base that is indicated as intronic (yellow), and an "N" for each base that lies outside of all the alignments (gray). Regions that are indicated as intronic by some alignments and exonic by others are also labeled "N".

3.2 ESTseq models, TWINSCAN_EST and N-SCAN_EST

The EST sequence can be exploited by any HMM-based gene predictor. Each state of the HMM is required to emit both the ESTseq and the target genome sequence from each state. When TWINSCAN uses ESTseq it emits ESTseq symbols, target genome bases, and conservation sequence symbols. Similarly, N-SCAN (Brown, Gross et al. 2005; Gross and Brent 2005) emits ESTseq symbols together with columns of multi-genome alignments. All states must have probability models for the emission of ESTseq symbols, so that these symbols can influence the likelihoods of functional annotations such as splice donor and acceptor, exon, intron, translation initiation and termination site, and so on. For example, the likelihood of emitting the I symbol from intron states should be greater than the likelihood of emitting I from exon states. Parameters for these models are estimated from examples of known gene structures together with their ESTseqs. Homogeneous Markov chains are used for ESTseqs in

UTR, intron region, intergenic region, and coding regions, and WAM (also called PSSM) in donor and acceptor sites. For TWINSCAN_EST on *C. elegans*, 1^{st}-order Markov chains were used for coding, UTR, intron states, and the translation initiation and termination signals. A 43-base long 2^{nd}-order WAM was used for acceptor splice site signals and 9-base long 2^{nd}-order WAM was used for donor splice site signals. Regions between 1000 bases and 150 bases upstream of the start of translation and downstream of the stop of translation were used as intergenic region. Intergenic regions' ESTseqs were used as the null model for each state. For TWINSCAN_EST on human, 5^{th}-order Markov chains were used for ESTseq in UTR, intron region, intergenic region and coding region.

For N-SCAN_EST on human, the single 5' UTR state is replaced by a 3' UTR state and four states associated with 5' UTRs. The four 5' UTR states model those a) unspliced UTRs from transcription start site (TSS) to the translation start site; b) initial noncoding exons (from the TSS to the splice donor site); c) internal noncoding exons (from acceptor site to donor site) and d) the noncoding segment of the exon containing the start codon (see Brown, Gross et al. 2005 for details). Each state will emit the genomic sequence, alignment sequences and ESTseq at each time. 5th-order Markov models were used for all ESTseq models except for the acceptor and donor splice site models, which were as for worm.

For the worm genome, when 5th-order Markov chains are used instead of 1st-order, the difference in accuracy statistics is not more than a fraction of one percent.

3.3 ESTseq Model Parameter Estimation

Given ESTseqs and their gene structure annotations, distinct sets of parameters for the models described in the previous section are estimated. For training and evaluation purpose, human RefSeq mRNAs excluding the predicted XM_ accessions, (Maglott, Katz et al. 2000; Pruitt, Katz et al. 2000; Pruitt and Maglott 2001) aligned to human genome (build NCBI35) were downloaded from the UCSC genome browser (http://genome.ucsc.edu). The RefSeq annotation was then cleaned up by removing genes with in-frame stop codons. There were 17,798 transcripts remaining, 17,120 of which contain UTR annotations. In order to estimate the ESTseq parameters, single-gene ESTseqs were cut out from the whole chromosome ESTseq with an additional 1000 bases on each end as the intergenic regions. The intergenic region model was used as the null-model. Parameters were estimated from these single-gene ESTseqs and their annotations.

3.4 Measuring Prediction Accuracy

Measuring prediction accuracy requires measures indicating the accuracy and a test data set with its genes annotated. For a fair comparison, the same accuracy measures and the same test dataset with a particular annotation set should be applied to all the methods to be compared. It is impossible to compare the accuracy of two prediction methods without a common test sequence dataset and a particular annotation set on the same set. Both the accuracy measurements and test data sets have evolved significantly in the past decade.

The most popular gene prediction accuracy measurement now is the three-level prediction accuracy measurement introduced by (Burset and Guigo 1996). The three levels are nucleotide level, exon level and protein product level. For each level, sensitivity (Sn) and specificity (Sp) are calculated to represent the accuracy. The meaning of sensitivity and specificity might vary from literature to literature. Here, the sensitivity is defined as the percentage of total annotations that are predicted correctly. The specificity is the percentage of overall predictions that are correct. For example, in nucleotide level, sensitivity and specificity can be defined as

$$Sn = \frac{correctly \quad predicted \quad coding \quad bases}{total \quad annotatted \quad coding \quad bases}$$

and

$$Sp = \frac{correctly \quad predicted \quad coding \quad bases}{total \quad predicted \quad coding \quad bases}.$$

On the exon level, sensitivity and specificity are defined similarly to those for nucleotide level except that a correctly predicted exon means every base of the exon is predicted exactly including the exon boundaries. Exon level sensitivity and specificity also show the accuracy of predictions on important signals, such as splice sites.

The protein product level can be further divided into transcript level and gene level because of alternative splicing. On transcript level, a transcript is predicted correctly if and only if all the coding region exons it contains are predicted exactly. On the gene level, a prediction is correct if at least one transcript from the same gene locus is predicted exactly. The protein product level measures not only how well the boundaries of coding regions and splice sites are predicted, but also how well the predicted exons can be assembled into final protein products, which include the start and stop of translation and exon frames.

For the first generations of gene prediction systems, it was very hard to generate a test data set with reliable annotations. A lot of effort was devoted to developing test data sets and their annotations. Many different test sets were constructed to evaluate the accuracy of gene prediction programs. The most widely used one among them might be the “Burset-Guigo” set (Burset and Guigo 1996), which contained 570 vertebrate genome sequences with one single gene in each sequence. The total size of the data set was less than 3 Mb and about 15% of the total bases therein were coding bases. It was far away from the reality. In a real human genome, there are only about 1% of the total bases used for coding, according to current estimation. More realistic test data sets were created in order to evaluate the gene prediction programs more accurately, like the “68-set” by Korf et al (Korf, Flicek et al.). It contained 68 contiguous mouse genome sequences with an average length of 112 Kb. With the progress of annotation projects like RefSeq and the finishing of human genome sequence, the gene prediction accuracy was finally evaluated on the entire human genome. Unfortunately, the more realistic the test set becomes, the lower accuracy gene predictors get. On the whole human genome, GENSCAN, the most accurate gene finder for human genome before the appearance of comparative-genomics-based methods, could predict only about 10% on gene level sensitivity (Flicek, Keibler et al. 2003), while TWINSCAN was reported in the same paper with gene level sensitivity 15%.

3.5 Accuracy of TWINSCAN_EST and N-SCAN_EST

The approach of using EST alignments for de novo gene prediction program described in the previous sections has been tested on TWINSCAN and N-SCAN. The new programs are called TWINSCAN_EST and N-SCAN_EST. Before using ESTs for gene predictions, with another worm genome (*C. briggsae*) available, the author had made TWINSCAN the best de novo gene predictor for worm genomes (Stein, Bao et al. 2003; Wei, Lamesch et al. 2005). With EST alignments, the accuracy of TWINSCAN_EST has been tested on two worm data sets (see section 3.5.1). N-SCAN had become the best de novo gene predictor for the human genome by using mouse, rat and chicken genomes as the informant without using any EST-alignment information (Brown, Gross et al. 2005). N-SCAN_EST has been evaluated on the whole human genome. Its novel predictions on human genome (build NCBI35) have been validated experimentally.

3.5.1 Data Sets

TWINSCAN_EST has been tested on two worm data sets. The first is the whole *C. elegans* genome (version WS130). *C. briggsae* version cb25.agp8 is used as the informant database. The *C. elegans* genome sequence version WS130 was downloaded from the WormBase website (Stein, Sternberg et al. 2001; Harris, Lee et al. 2003; Harris, Chen et al. 2004). The *C. briggsae* genome sequence version cb25.agp8 was downloaded from the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/wormbase/cbriggsae/cb25.agp8). In total 302,075 *C.*

elegans ESTs were downloaded from dbEST (Boguski, Lowe et al. 1993) (http://www.ncbi.nlm.nih.gov/dbEST/) (1/20/2005 version).

The second set is the GAZE data set, 2 Mb created by concatenating the sequences of 325 genes flanked by half the intergenic region to the closest known gene on each side (Howe, Chothia et al. 2002). The genome sequence for the GAZE data set was downloaded from http://www.sanger.ac.uk/Software/analysis/GAZE. The informant database (*C. briggsae*) was as above.

Human ESTs were downloaded from dbEST on January 20th 2005. The informant database for TWINSCAN is mouse genome (Waterston, Lindblad-Toh et al. 2002) Build 33 (mm5 on the UCSC browser). Other informant datasets for N-SCAN include mouse, rat (Gibbs, Weinstock et al. 2004) (UCSC rn3) and chicken (Hillier, Miller et al. 2004) (UCSC Galgal2) genomes (Brown, Gross et al. 2005; Gross and Brent 2005).

3.5.2 Genome Alignments

For worm data sets, conservation sequences were generated from WU-BLAST (http://blast.wustl.edu) alignments of the whole *C. elegans* genome against *C. briggsae* genome. First, *C. briggsae* sequences longer than 150kb were cut into 150kb sequences with 20kb overlap, and then the Blast database was generated from all sequences after they had been masked by NSEG with default parameters. BLASTN parameters were: M=1 N=-1 Q=5 R=1 B=10000 V=100 lcfilter filter=seg filter=dust topcomboN=1.

For the human data set, an informant database from mouse genome was constructed for TWINSCAN in a similar way as the informant database from *C. briggsae* for *C.elegans*. The human chromosomes were split into 1Mb fragments first, and then conservation sequences for TWINSCAN were constructed for each fragment. The alignment sequences for N-SCAN were constructed for each fragment by using the

BLASTZ alignments of human, mouse, rat and chicken genomes (Gross and Brent 2005).

3.5.3 ESTseqs

C. elegans ESTs were aligned to WS130 by using standalone version 25 of BLAT (Kent 2002). ESTseqs were generated using only those EST alignments in which the number of matches was at least 95% of the length of the entire EST, including unaligned portions. These alignments were projected onto a genomic sequence to generate its ESTseq. Similar procedures were performed for the GAZE sequence.

Human ESTs were aligned to the whole human genome by BLAT. Only the EST alignments in which the number of matches was at least 98% of the length of the entire EST were chosen to project to the genomic sequence to generate ESTseqs. The ESTseq of each chromosome was then split into 1Mb fragments corresponding to the 1Mb fragments of genomic sequences.

3.5.4 Accuracy Evaluation: *C. elegans*

For the whole genome worm data set (WS130), TWINSCAN_EST's performance was tested by 8-fold cross validation. The whole genome was split into fragments of about 500kb. Each fragment was assigned to one of eight groups randomly. TWINSCAN_EST was trained on fully confirmed genes from seven of the eight groups, and run on the fragments from the eighth group to avoid training and testing on the same data set. The results show 14% improvement in sensitivity and 13% in specificity in predicting exact gene structures compared to pure TWINSCAN 2.03 (Figure 3-2). TWINSCAN 2.03 was, in turn, significantly more accurate than both FGENESH (v.1, with *C. elegans* parameters v.1) (Salamov and Solovyev 2000; Solovyev 2002) and GENEFINDER (release 980504, P. Green, unpublished), the two most widely used prediction programs for nematodes.

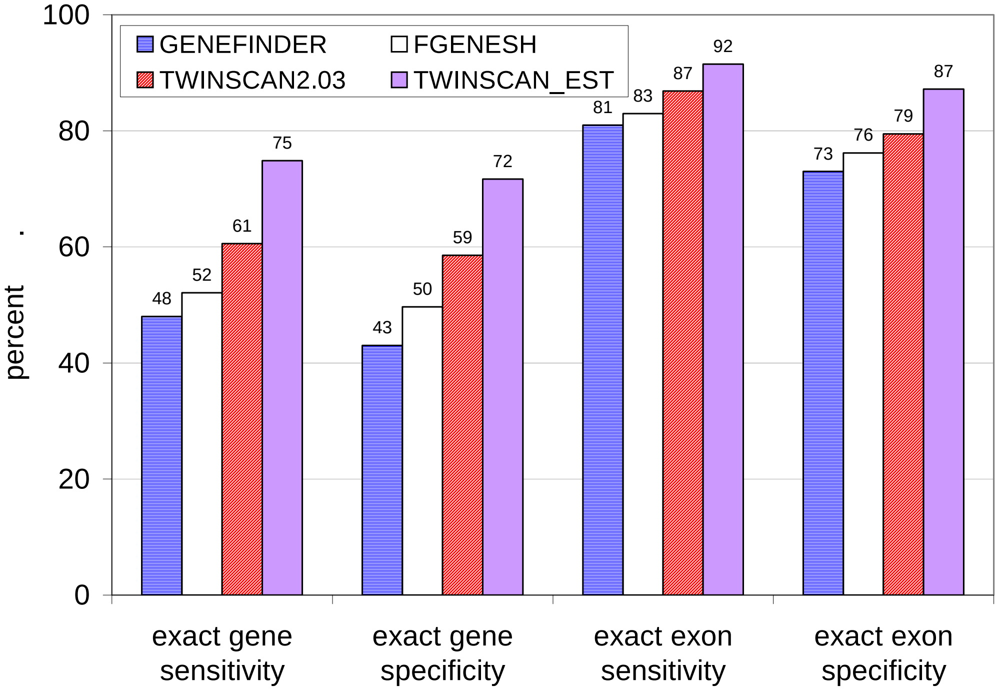


Figure 3-2. Results on the whole C. elegans genome.

C. briggsae was used as the informant database and C. elegans ESTs were from dbEST. The sensitivities are based on the 4,705 fully confirmed genes from WS130 and the specificities are based on those predictions that overlap with fully confirmed genes.

The second test used the 2 Mb GAZE dataset, which was created by concatenating the sequences of 325 genes flanked by half the intergenic region to the closest known gene on each side (Howe, Chothia et al. 2002). For TWINSCAN_EST on GAZE data set, no cross validation was applied, parameters were estimated by author from all fully confirmed genes of WS130. Since the number of the fully confirmed genes of WS130 (about 5000 genes) is much larger than 325, the estimated parameters are unlikely to be biased on the 325 genes.

In order to show that the improvement on performance was not caused by the EST-alignment-selection procedure, the exact same EST alignments were used for both TWINSCAN_EST and GAZE_est on the GAZE data set. *C. elegans* ESTs were downloaded from dbEST (1/20/2005), aligned to the GAZE genomic sequence by using BLAT, and filtered for alignment quality (the same as for the whole C. elegans genome). The remaining EST alignments were used for both TWINSCAN_EST and GAZE_est. The results show that TWINSCAN_EST is more accurate than GAZE_est, especially for exact gene structure prediction. TWINSCAN_EST has 73% gene sensitivity and 62% gene specificity compared to GAZE_est's 61% and 58% (Figure 3-3). This experiment demonstrated the efficiency of the novel approach to exploit the EST alignments for gene prediction.

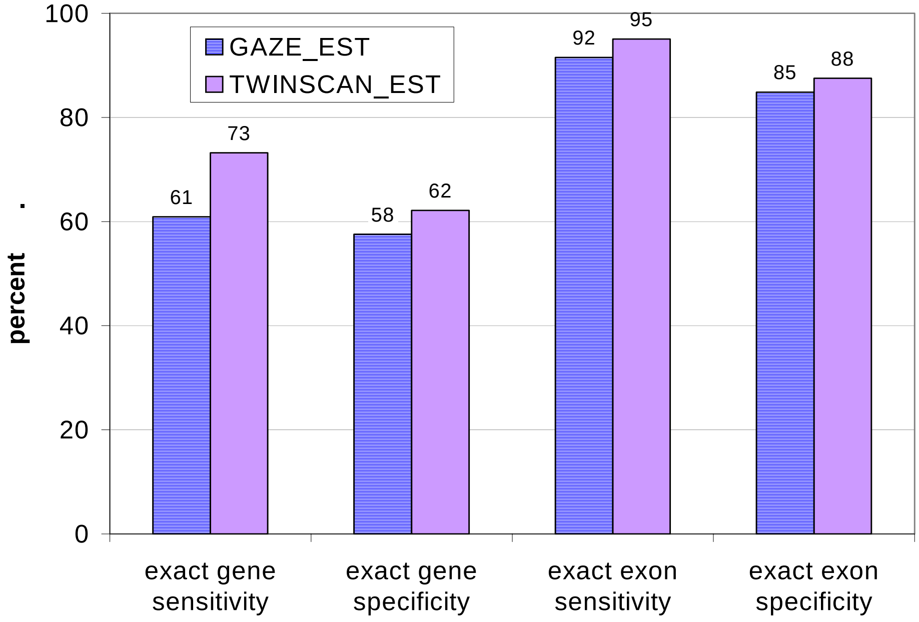


Figure 3-3. Result on GAZE merged data set.

Both TWINSCAN_EST and GAZE_EST used the same EST alignments from dbEST (1/20/2005). 305 of the 325 gene loci have at least one EST alignment. *C. briggsae* was used as the informant genome.

Although TWINSCAN_EST shows substantial improvement compared to previous systems based on fully confirmed worm genes, these genes are more likely to have aligned ESTs than a randomly selected gene. Thus, an independent test is needed in order to determine how TWINSCAN_EST would perform on genes with no aligned ESTs. Such a test was performed by running TWINSCAN_EST on the entire genome with an empty EST database, so that no gene had aligned ESTs. This resulted in slight improvements to sensitivity and specificity in exact gene prediction compared to predictions by TWINSCAN 2.03, which does not consider the presence or absence of ESTs (Table 3-1). These improvements may result from applying a slight score penalty to exons and genes without ESTs – in this case all exons and genes. Such a penalty would eliminate predicted exons and genes with marginal scores, in effect filtering out the lowest scoring predictions from TWINSCAN 2.03. Since the lowest scoring predictions are mostly incorrect, this would improve accuracy. On the other hand, the improvement in gene accuracy is small, and exon accuracy does not improve, so it is safe to conclude that novel genes with no ESTs are predicted with approximately the same accuracy by TWINSCAN_EST and TWINSCAN 2.03.

Table 3-1. Result for deletion experiment.

The first column is for TWINSCAN2.03 and the remaining 3 columns are for TWINSCAN_EST. The second column is for the TWINSCAN_EST performance with empty ESTseq, i.e. all bases in ESTseqs are 'N's. For the third and fourth column, 10% of genes in the annotation were set to "N"s. The third column is for TWINSCAN_EST's performance on the 10% genes with masked ESTseqs and the last column is for the 90% genes with unmasked ESTseqs. Results show that EST alignments improve the prediction accuracy but do not compromise the capability to predicted novel genes where EST alignments do not exist (column 2). Specificities are based on predictions that overlap with annotations by at least 1bp.

	TWINSCAN2.03	TWINSCAN_EST		
		Blank ESTseq	10% with ESTseq masked	90% with ESTseq unmasked
Gene_sn	60.6	61.3	63.0	74.7
Gene_sp	58.6	59.8	60.5	71.5
Exon_sn	86.9	86.2	86.4	91.5
Exon_sp	79.5	80.8	81.1	87.0

The previous experiment in which all ESTs were deleted from the database may yield an overly pessimistic assessment of TWINSCAN_EST's accuracy on novel genes with no aligned ESTs. It is possible that the presence of EST alignments for some genes may improve the accuracy of TWINSCAN_EST on the neighboring genes even when those neighboring genes have no aligned ESTs. The intuition is that certain kinds of mistakes, such as incorrectly splitting a gene with an EST and joining part of it to a neighbor without an EST will become much less common. To test whether such indirect benefits actually exist, we did a partial EST deletion experiment. All fully confirmed WS130 genes were divided into 10 groups at random, each containing 10% of the fully confirmed genes. One group of fully confirmed genes was selected, its ESTseq were masked with "N", and TWINSCAN_EST was run on the entire genome. These steps were repeated 10 times. Each time, the ESTseq for a different 10% of the confirmed genes was masked, so that the ESTseq for each confirmed genes was masked in exactly one repetition. We then computed the average accuracy statistics over the 10 runs for both the masked and unmasked genes. Results are shown in Table 3-1. The gene sensitivity of TWINSCAN_EST on the genes with masked ESTseq was 2.4% higher than TWINSCAN 2.03 and the specificity was 1.9% higher. In addition, exon and gene accuracy were higher than TWINSCAN_EST with blank EST sequence, indicating that the presence of ESTs for other genes did indeed improve the accuracy of genes with no ESTs.

The previous experiments showed TWINSCAN_EST's accuracy on genes with or without aligned ESTs. In practice, there are many genes covered partially by ESTs. To investigate the effect of partial EST coverage, we did the following experiment. ESTseqs were generated as in TWINSCAN_EST experiment for Table 3-1. For each fully confirmed WS130 gene, 50% of its gene region of the ESTseq was randomly masked as following: Let [0, 1] stand for a gene region. A random number *a* in range [0, 1] is generated, and then all bases in region [*a*, *a*+0.5] were masked with "N" if *a*<=0.5 or bases in region [*a*, 1] U [0, *a*-0.5] were masked with "N" if *a*>0.5, so at least 50% of each gene region was not covered with any EST alignment.

TWINSCAN_EST was then run on the entire genome with ESTseqs generated as above. The predictions were evaluated on all the confirmed genes. The gene sensitivity is 69%, which is about halfway between the gene sensitivity of pure TWINSCAN 2.03 (61%) and TWINSCAN_EST without ESTseq masking (75%). The gene specificity is 67%, which is about two-thirds of the way between that of pure TWINSCAN 2.03 (59%) and TWINSCAN_EST without ESTseq masking (71%).

3.5.5 Accuracy Evaluation: Human

TWINSCAN_EST and N-SCAN_EST on the whole human genome

TWINSCAN_EST and N-SCAN_EST were also tested on the whole human genome (build NCBI35) (Figure 3-4). TWINSCAN_EST on this data set produced about 10% improvement in sensitivity and 3% in specificity in predicting exact gene structures compared to pure TWINSCAN 2.03. N-SCAN_EST on NCBI35 produced a 6% improvement in sensitivity and 1% in specificity on exact gene structure level compared to N-SCAN. The relative less improvement from N-SCAN to N-SCAN_EST might because that N-SCAN_EST is built upon N-SCAN, which has higher accuracy on human genome than TWINSCAN does. Therefore some newly predicted genes by TWINSCAN_EST were correctly predicted by N-SCAN already.

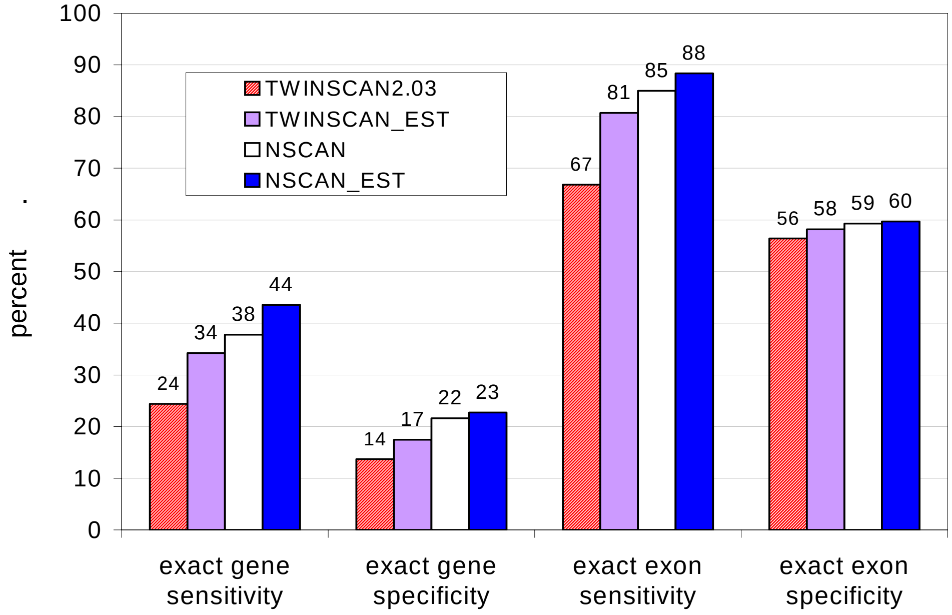


Figure 3-4. Comparison of TWINSCAN, TWINSCAN_EST, NSCAN and N-SCAN_EST results on human genome.

For TWINSCAN and TWINSCAN_EST, Mouse genome is used as the informant database. For NSCAN and N-SCAN_EST, mouse, rat and chicken genomes are used as the informant databases. Human ESTs are from dbEST. For all methods, pseudo genes are masked out first (van Baren and Brent 2005). Specificities are lower than sensitivities because they are based on all predicted genes, not only those that overlap known genes.

N-SCAN_EST and AUGUSTUS+ Comparison on Human Chromosome 22

AUGUSTUS+ is a newly reported gene prediction system which also explicitly evaluated the effect of EST alignments on gene prediction. In order to do a fair comparison to AUGUSTUS+, BLAT alignments of all spliced human ESTs on human chromosome 22 (version hg17) were downloaded from the spliced human EST track in UCSC genome browser (http://www.genome.ucsc.edu) on March 12th, 2006. Both

AUGUSTUS+ and N-SCAN_EST started from these EST alignments. EST parameters for N-SCAN_EST were estimated from the cleaned RefSeq annotations on chromosome 1, 2, 20 and 21. Parameters for AUGUSTUS+ EST hints were estimated by its author from chromosome 21. Comparing the results to RefSeq genes on chromosome 22, N-SCAN_EST's sensitivity and specificity on gene level were 47.1% and 23.6%, respectively. The comparable numbers for AUGUSTUS+ using the same EST alignments were 37.9% and 19.4%, respectively. N-SCAN_EST's result is significantly better than AUGUSTUS'. Part of the reason is that N-SCAN_EST is built upon N-SCAN, which is better than most of de novo gene predictors, including AUGUSTUS. Unlike AUGUSTUS+, which only uses the ESTs verifying a coding region structures, every aligned EST bases, no matter in a coding region or in a UTR region, are used by N-SCAN_EST. The result above also demonstrates that this novel method of using EST alignments the author developed here is effective even for gene prediction systems like N-SCAN, which is already very accurate.

3.6 Experimental Validation of N-SCAN_EST Novel Predictions on the Human Genome

Although the ultimate goal of gene structure prediction is to build a system that is accurate enough so that experimental verification and manual curation is no longer necessary, all the gene structures predicted computationally now still need to be verified experimentally. The novel introns predicted by N-SCAN_EST were selected (see Section 3.6.1), and tested by performing RT-PCR, sequencing and aligning the resulting sequence back to the whole human genome.

3.6.1 Target Selection

Introns were defined as novel if at least one of their splice sites was not in a region previously known to be transcribed – that is, not in an intron or exon defined by the alignment of any human mRNA or the spliced alignment of any EST or RefSeqs in which matching bases make up of at least 95% of the length of the expressed sequence. For this experiment, novel introns were picked from predicted genes of which another part was in a region known to be transcribed. N-SCAN_EST predictions (23,265 genes) were divided into three categories: fully within regions known to be transcribed, partially within regions known to be transcribed, and completely novel. The middle category, called partially novel, was divided into those with at least one novel splice site and those without any novel spliced site. For those partial novel genes with at least one novel splice site, we filtered out predictions containing processed pseudogenes in their exons using the method described in (Van Baren and Brent 2005). The remaining predictions with at least one novel splice site were used for primer design.

3.6.2 Primer Design

Primers were designed to amplify approximately 800bp of cDNA spanning at least one targeted intron by using Primer3 (Rozen and Skaletsky 2000; http://frodo.wu.mit.edu/cgi-bin/primer3/primer3_www.cgo) with default parameters except for PRIMER_MIN_SIZE = 17, PRIMER_MIN_GC = 30, PRIMER_MAX_GC = 70, PRIMER_OPT_TM = 70, PRIMER_MIN_TM = 65, PRIMER_MAX_TM = 75, and PRIMER_GC_CLAMP = 2. Amplicons were designed to be as long as possible without exceeding 800 nt. By this, primer pairs for 748 partial novel predictions were generated successfully.

3.6.3 PCR and Sequencing

PCR amplification and sequencing were as described in (Brown, Gross et al. 2005).

3.6.4 Sequencing Result Analysis

The RT-PCR and sequencing results were aligned to the whole human genome by using BLAT (standalone version 25). BLAT parameters were "-repeats=lower -ooc=11.ooc -q=rna". When an experimental sequence could be aligned to multiple loci, its genomic locus was determined by the BLAT alignment containing the greatest number of matching bases. The genomic region was cut out with an extra 1000 bases on both sides. EST_GENOME was then used to fine tune the alignment. Sequences lacking 50 consecutive matches were classified as "bad". The remaining sequences were classified as "hit" if their best alignment met the following criteria:

1) At least 95% of the entire sequence were matches;

2) It overlapped the target gene region;

3) It contained at least one intron; and

4) All 10-bp sequences flanking introns contained at least 8 matches.

The target intron was considered "verified" if the best alignment contained an intron that exactly matched it.

150 gene predictions yielded an experimental sequence alignment that (1) had at least one intron, and (2) overlapped the targeted gene. Among these, there were 63 in which the targeted intron was confirmed exactly at both splice sites. The remaining 87 gene predictions included cases where the boundaries of the target intron could not be determined because the high quality portion of the sequencing read or reads did not extend that far. Since primer pairs were designed to amplify approximately 800bp of cDNA spanning the target intron, the alignments of many amplicon sequences determined more than one novel intron. In total, 332 novel introns were

experimentally determined, of which 127 matched predicted introns exactly at both splice sites.

All sequences have been submitted to dbEST (http://www.ncbi.nlm.nih.gov/dbEST) and traces to the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi). Accession numbers are provided in appendix B.

The experimental verification procedure has been improved since the above mentioned experiments. By changing the cDNA spanning length from 800 bases to 500 bases, the specificity of verified splice sites was improved since the resulted sequences are less likely to be from a region amplified before. Up to date, RT-PCR experiments targeting predictions from N-SCAN_EST have led to verification of more than a thousand completely novel human introns and their surrounding exons. This is a significant achievement especially for a well studied genome like the human genome.

3.6.5 An N-SCAN_EST Prediction Example

Figure 3-5 shows an N-SCAN_EST prediction with novel introns verified by the EST sequence alignments. On the top of the figure are coordinates in human genome chromosome 3. GENSCAN predicted two separate genes in this region while N-SCAN_EST only predicted one gene. Ensembl did not predict any gene in this region since there were no existing mRNAs or proteins similar to this region then. The ESTs available then were also not enough to generate an Ensembl prediction in this region. Geneid and SGP predicted different structures in the first intron N-SCAN_EST predicted. ESTs DT932676 and DT932675 were derived from N-SCAN_EST predictions followed with PCR amplification and sequencing. All introns predicted by N-SCAN_EST were verified by EST DT932675. Although an independent human mRNA DQ232881 and its consequence RefSeq gene were generated several months after the ESTs from N-SCAN_EST predictions had been submitted to dbEST, the alignment of this human mRNA confirmed the full-length coding region predicted by N-SCAN_EST.

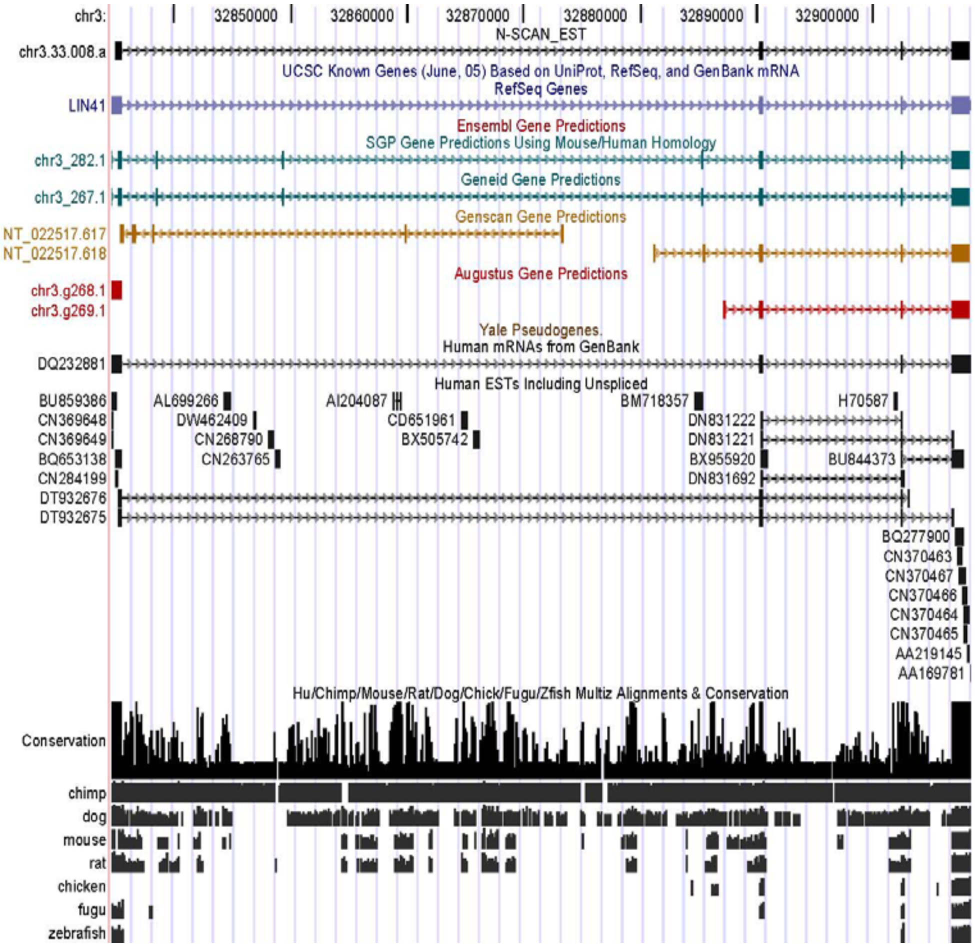


Figure 3-5. An N-SCAN_EST partial novel prediction example.

On the top of the figure are coordinates in human genome chromosome 3. From top to bottom, the first track is for N-SCAN_EST and the following two tracks are for UCSC known genes, RefSeq genes. The next several gene prediction tracks are for Ensembl, SGP, Geneid, GENSCAN and AUGUSTUS gene predictions respectively. Human mRNA and EST tracks are under the pseudogene track, which shows no pseudogene here. The bottom tracks are for the Human ESTs followed evolutionary conservation derived from Multiz alignments. The genomes and their assemblies used in Multiz alignments were human (hg17), chimp (panTro1), mouse (mm5), rat (4n3), dog (canFam1) chicken (galGal2), fugu (fr1) and zebrafish (danRer1).

3.7 Effect of EST Coverage

For well-studied model organisms, such human and *C. elegans*, their ESTs might cover more than 80% of the total genes, while for newly sequenced genomes like chimp or dog, there are few ESTs for them. In order to investigate the trainability of the model and the effect of EST coverage, *C. elegans* genome sequence version WS130 was downloaded from WormBase and *C. briggsae* was used as the informant genome as in section 3.5. 10%, 20%, 50% and 100% of all *C. elegans* ESTs from dbEST (date 1/20/2005) were used to generate ESTseqs for WS130 to represent the different coverage. All 500-kb fragments of WS130 sequences were divided into 2 groups. Known genes inside one group were used as the training set, and parameters estimated from this training set were tested on all fragments in the other group. A parameter file was estimated from the training set for each EST coverage rate. Each parameter file was then used to test on the test sets with ESTseqs derived from all ESTs. These steps were repeated by reversing the training and testing groups. Then the prediction results with the parameter files estimated from different levels of coverage rates were evaluated on the overall confirmed gene set. The gene level sensitivity results are shown in Figure 3-6. The results indicate that the parameter file trained from higher EST coverage has better accuracy. The same trend holds for gene level specificity too.

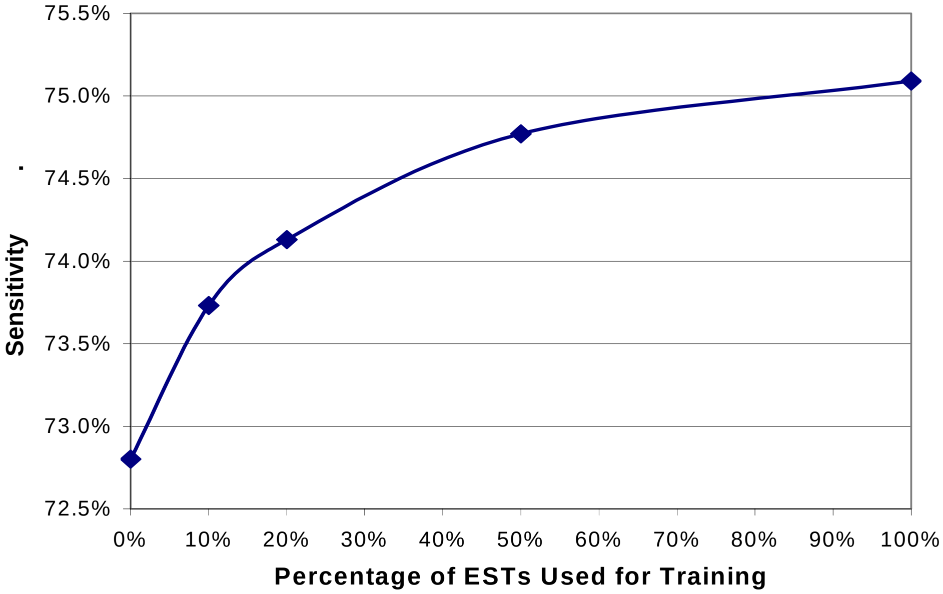


Figure 3-6. Gene-level sensitivity versus EST coverage used for training.

Along x axis, from left to right, are four sets of ESTseqs derived from 10, 20, 50 and 100% of the overall C. elegans ESTs in dbEST 1/20/2005. Genomic sequences and informant database are the same as in other TWINSCAN_EST experiments on WS130. Parameters estimated from these ESTseqs using the confirmed C. elegans gene set. From left to right, the nodes are for results using parameters estimated from ESTseqs with 10%, 20%, 50% and 100% of ESTs in dbEST 1/20/2005. 2-fold cross-validation was used for all experiments.

3.8 Trainability of TWINSCAN_EST

All experiments shown above used ESTseq parameters estimated from the alignments of ESTs to genomic sequences from the same organism. For newly sequenced genomes, there might be not enough annotated genes and ESTs to train the parameters. How much can we gain by using the ESTseq parameters estimated from the native EST-to-genome alignments, i.e., the EST and genome sequences of the alignments are from the same organism? In order to investigate this, ESTseq parameters were estimated from human dataset, which includes the human RefSeq annotations and alignments of human ESTs from dbEST (1/20/2005) against human genome. All 1Mb fragments of human genome were divided into 2 groups randomly. RefSeq genes contained inside any fragments of one group were used as the annotations to train the ESTseq parameters. Then TWINSCAN_EST was run on fragments belonging to the other group. The procedure repeated twice for human-by-human experiments. One set of the retrained human ESTseq parameters were combined to TWINSCAN parameter for worm, after which TWINSCAN_EST was run for all the *C. elegans* fragments. Similar procedure was done for *C. elegans* dataset. Here ESTseq models for both human and worm are the same, which was 5th-order Markov models for coding regions, UTRs, introns, intergenic regions and 43-base-long 2nd -order WAM for splice acceptor site and 9-base 2nd order WAM for splice donor site. The results in Table 3-2 show that the performance of TWINSCAN_EST can be improved modestly but clearly by using parameters derived from the ESTs and genomic sequences from the same organism.

Table 3-2. TWINSCAN_EST performance when ESTseq parameters were trained from the same or different organism.

Gene and exon level sensitivity and specificity and the average of sensitivity and specificity are shown for each case. From the top to bottom, the first row is for TWINSCAN_EST performance on human genome with ESTseq parameters estimated from alignments of human ESTs to human genome; the second row is for human genome with ESTseq parameters estimated from alignments of *C. elegans* ESTs to *C. elegans* genome. For genome sequence and the parameter file are from the same organism, 2-fold cross validation were used.

	G_sn %	G_sp %	(G_sn +G_sp)/2 %	Ex_sn %	Ex_sp %	(Ex_sn +Ex_sp)/2 %
Human_by_human	32.6	16.6	24.6	77.9	58.5	68.2
Human_by_worm	31.7	16.3	24	76.2	58.6	67.4
Worm_by_worm	75.4	71.9	73.7	91.9	87.5	89.7
Worm_by_human	74.5	70.6	72.6	91.9	86.9	89.4

3.9 Using ESTs from Different Organisms

The number of ESTs from the same organism for many genomes (like the rat genome) is not big enough to cover most of their genes yet, but a lot of times, many more ESTs can be found from closely-related species. No existing gene prediction methods have effectively used those EST data.

Aligning ESTs from different organisms will have alignment discrepancies due to sequencing error and evolutionary divergence. No existing spliced alignment program can deal with cross-species spliced alignment very well. The new spliced alignment algorithm described in Chapter 5 might be tuned for this purpose. Further discussions regarding to this may be found at the end of Chapter 5.

This section shows the gene prediction on rat genome using ESTs from mouse and/or human genome. Both human and mouse have millions of ESTs available, while fewer ESTs are available for rat. The default alignment program BLAT was still used here. Rat genome (version rnv3) was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/rnJun2003/bigZips/chromFa.zip. Human genome (build NCBI35) was used as the informant to generate conservation sequence for the rat genome. Human ESTs and mouse ESTs were from dbEST (1/20/2005). Rat ESTseqs were generated in three different ways, using mouse ESTs only, human ESTs only, and both. For all these three cases, only those alignments of which the number of matches was at least 95% of the entire EST length were chosen to generate ESTseqs. The results are shown in Table 3-3. It indicates that the closer evolution distance is between two species, the higher value will the EST-to-genome alignments contribute. In general, however, cross-species EST alignments generated much less improvement compared to the improvement derived from native EST-to-genome alignment, such as the improvement in human gene prediction from human ESTs aligned to human genome. One reason why using human ESTs produces less improvement is that the number of ESTs actually used to generate the ESTseq is much less since we used a relative high alignment quality threshold (95%) for both mouse ESTs and human

ESTs. If we lower the threshold for human ESTs, the evolutionary divergences between the two genomes will make it very difficult to align the ESTs accurately to the right positions. A more accurate cross species alignment program might be helpful here.

Table 3-3. TWINSCAN_EST using cross-species EST-genome alignments.

TWINSCAN_EST was run on rat genome (version Rnv3) with human genome as the informant for conservation sequence and ESTseqs derived from mouse and/or human ESTs. Only those EST-genome alignments for which the number of matches was at least 95% of the length of the entire EST were chosen to generate ESTseqs. From the top to the bottom, TWINSCAN result is shown in the first line; TWINSCAN_EST result using mouse ESTs only is in the second line; and the result in the third line used human ESTs only. The last line stands for TWINSCAN_EST using both human and mouse ESTs. ESTseq parameters were estimated from the 3,405 rat RefSeq genes and a 2-fold cross validation was used for each of the three experiments using ESTseqs. The specificity is much lower than that on human genome since the total number of transcripts in the rat annotation set is only about one sixth of that of the human genome.

	Gene_sn %	Gene_sp %	Exon_sn %	Exon_sp %
TWINSCAN	21.9	2.7	70.3	13.6
TS_EST mouse ESTs	23.7	2.9	73.1	13.9
TS_EST human ESTs	22.2	2.7	70.8	13.7
TS_EST mouse+ human ESTs	23.8	2.9	73.1	13.9

3.10 Effect of Different EST-to-genome Alignment Programs

Different EST-to-genome alignment algorithms such as BLAT and EST_GENOME have been tested as the underlying alignment program for the gene prediction systems using ESTseqs. EST_GENOME uses dynamic programming to find the optimal spliced alignment of a cDNA and a genomic sequence. It is a slow but accurate program while BLAT is fast but not as accurate as EST_GENOME (Kent 2002). BLAT results are not highly reliable either, though in C. elegans data sets, the accuracy difference between the TWINSCAN_EST systems using BLAT and EST_GENOME is not very big. In rice genome, we did observe that the alignment accuracy affected the gene prediction result greatly since the splicing site structure of rice genome crashed BLAT a lot of times. The effect of different alignment programs will be further discussed in Chapter 5.

3.11 Using EST Alignments for 5'UTR Prediction

Both N-SCAN and N-SCAN_EST can predict complete gene structures with 5'UTRs. Since ESTs include UTR regions and coding regions, better accuracy of UTR prediction is expected with EST alignment information. Table 3-4 shows the accuracy of 5'UTR predictions by N-SCAN and N-SCAN_EST on human genome NCBI35 shown in section 3.5.5. The 5'UTRs in the RefSeq annotation set generated in section 3.3 was used as the 5'UTR annotations which contained 21,675 5'UTRs. An initial exon is the first coding region exon at downstream of a 5'UTR. The total number of initial exons in the annotation set is 13,087. The prediction accuracy of initial exons is also presented in the table. With ESTseq model, the 5'UTR predictions are 3% more sensitive on exon level and the initial exon 5.5% more sensitive without decreasing the specificity. Since more ESTs cover 3' UTRs than 5' UTRs, the accuracy of 3'UTR prediction may benefit more from EST alignment information. However, no reliable 3'UTR annotation set has been created to verify this currently.

Table 3-4. EST Alignment Effect on 5'UTR and Initial Exon Prediction.

5'UTR predictions on exon level and initial exon prediction accuracy are reported here.

	5’UTR Sensitivity %	5’UTR Specificity %	Initial Exon Sensitivity %	Initial Exon Specificity %
N-SCAN	11.24	5.88	59.44	39.89
N-SCAN_EST	14.20	6.75	65.02	41.53

3.12 Another Example of N-SCAN_EST Prediction on the Human Genome

N-SCAN_EST was also independently evaluated on the human ENCODE regions as part of the recent EGASP community evaluation (Guigo and Reese 2005). 12 groups took part in the competition.

Figure 3-7 presents an example of N-SCAN_EST predictions. This example demonstrates the power of combining the comparative-genomics-based gene prediction methods and EST alignment information. There are high quality EST alignments in the region, such as BX116511 with a 100% identical alignment of size 583 bases. These EST alignments make N-SCAN_EST one of the only two gene predictors that predict any gene in this locus. The other gene predictor is AceView, which is a transcript-alignment-based gene prediction method using all publicly available ESTs and mRNAs as well. Three of the four annotated exons were predicted correctly by N-SCAN_EST. N-SCAN_EST missed an exon even though there was EST evidence for it. Further investigation in the region found that the conservation level at the region was very low and it overwhelmed the EST evidence eventually. This example shows that for a region with extreme low conservation level, N-

SCAN_EST might skip this region even with the appearance of an EST evidence. However, with EST evidence, the bias against relatively-new genes of a species, thus with a low conservation level, can be reduced. The most parts of the first exon and last exon in this predicted gene are also in low-conservation-level regions. By combining the power of comparative-genomics-based gene prediction and EST alignments, N-SCAN_EST could predict this gene while most of other gene predictors have missed it completely.

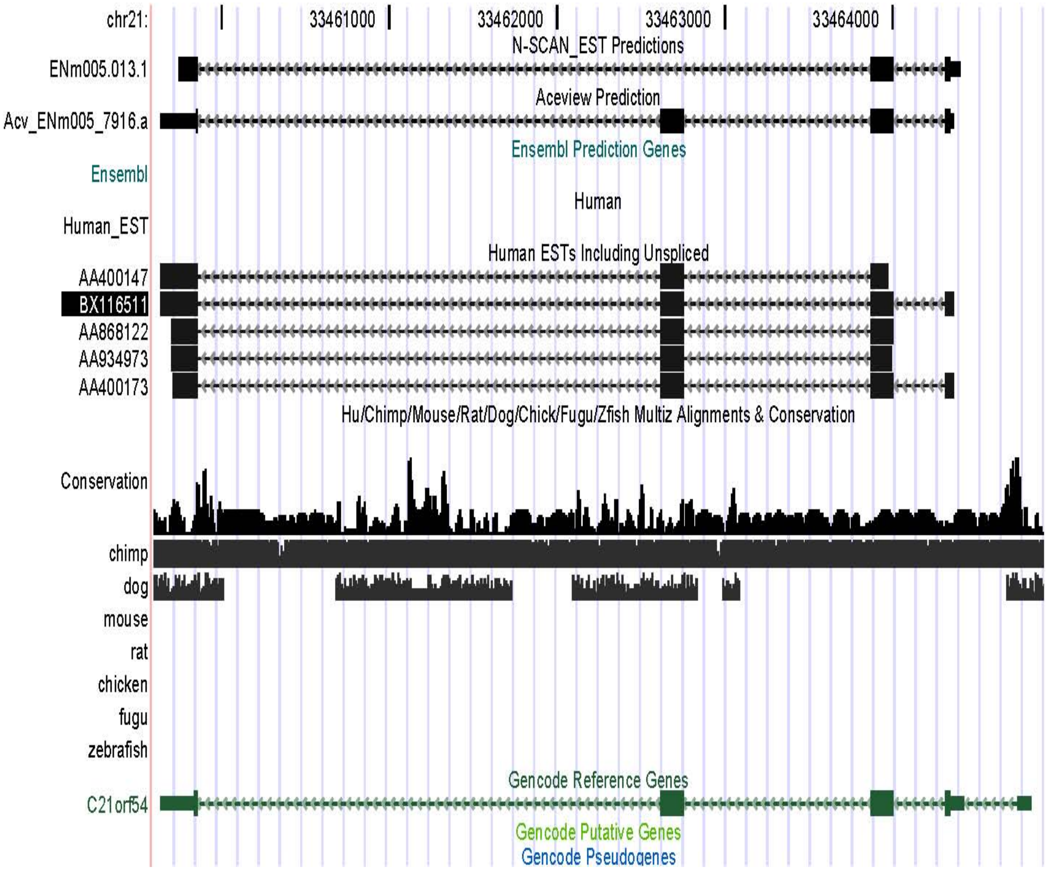


Figure 3-7. An example of N-SCAN_EST predicted genes.

From top to bottom, the first track is for N-SCAN_EST, the second track is for Aceview and the third track is for Ensembl which has no prediction here. The middle tracks are for the Human ESTs followed by Multiz Alignments and conservation level. The track at the bottom is the ENCODE annotation. N-SCAN_EST predicts two out of the four exons exactly, and another exon partially. N-SCAN_EST missed the second exon from left even though there were ESTs for the region. The conservation rate for this region is low, especially on the region near the second exon in the ENCODE annotation.

3.13 DISCUSSION

The method introduced in this chapter for integrating information from EST alignments and a HMM-based gene predictor has five key features:

1) It combines the power of comparative-genomics-based gene prediction programs and EST alignments;

2) It significantly improves the gene prediction accuracy for genes with EST alignments;

3) The accuracy of gene prediction for genes without EST alignments is at least as good as that of the original gene prediction system without considering EST alignments;

4) The accuracy of gene prediction is improved for genes without EST alignments if these genes are interspersed with genes that have EST alignments;

5) ESTs regions aligned to UTR region can improve the UTR region prediction as well as the coding region prediction. Therefore, every aligned EST bases contribute to gene prediction.

Therefore, the use of EST information is very effective and comes at no cost. TWINSCAN_EST and N-SCAN_EST have the essential benefit of a de novo gene finder, the ability to find completely novel genes without sequence similarity to known genes. And they are more accurate on genes for which ESTs are available. Compared to other de novo gene finders, TWINSCAN_EST is the most accurate program available for nematodes and N-SCAN_EST is the most accurate program available for mammals. Thus, it is highly recommended to use the EST versions of these programs on any genome for which there is EST information.

These programs can also be valuable components of multi-stage “pipelines” for gene structure annotation. For example, a new gene annotation system has been built recently, in which the first stage is aligning full-ORF cDNA sequences to their native locus using a new cDNA-aligner, Pairagon (Arumugam et al., submitted, http://genes.cs.wustl.edu/BrentLab/MB-Lab-Software.html). Where there is no full-length cDNA to align, N-SCAN_EST predictions are used with ESTseq created from BLAT alignments of ESTs. This system was independently evaluated on the human ENCODE regions as part of the recent EGASP community evaluation (Guigo and Reese 2005). The results showed that the accuracy of this system was comparable to that of the ENSEMBL pipeline (slightly better by most measures), even though ENSEMBL makes use of non-human protein alignments in addition to human transcripts. Since native human cDNA and EST sequences may be considered preferable evidence compared to non-native alignments, Pairagon+N-SCAN_EST is an attractive option. It is also simple to use in that it does not need an elaborate list of heuristics to determine how the sequences from different sources should be aligned to each locus. Figure 3-8 showed exon-level performance of the gene annotation pipeline N-SCAN_EST + Pairagon and the pipeline using full-length cDNA Pairagon alignments only. It demonstrated that the fewer cDNAs there are the more N-SCAN_EST improves the performance. Since there are many genomes for which significant numbers of ESTs are produced but few if any high quality cDNA sequences are produced by sequencing for full-length cDNAs, the gene annotation accuracy can be improved significantly for those genomes.

Finally, it has been demonstrated that N-SCAN_EST is sufficiently specific that RT-PCR experiments designed to amplify its predictions can produce new significant numbers of cDNA sequences even in the extremely well-studied genome of Homo sapiens.

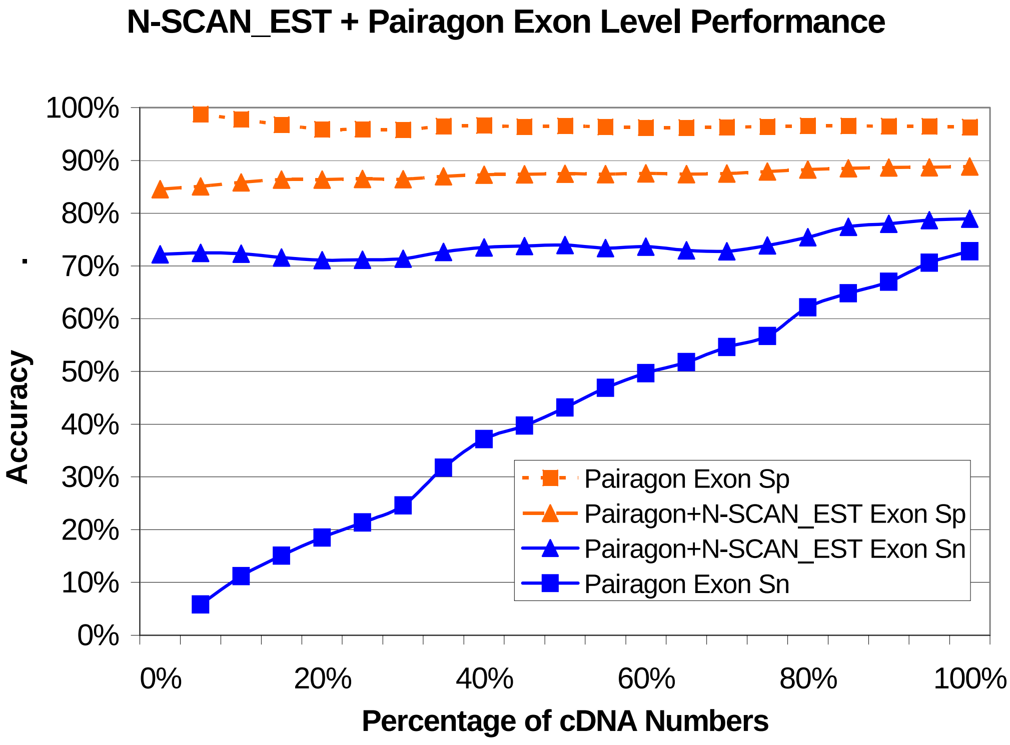


Figure 3-8. The exon-level performance comparison of N-SCAN_EST+Pairagon with different number of cDNAs available.

On the 31 ENCODE regions, there were 445 cDNAs aligned to them. The x axis stands for percentage of the 445 cDNAs included in result annotation. From left to right, 5% of cDNAs were randomly picked and added to Pairagon program and N-SCAN_EST+Pairagon. The curves with square dots stand the performance of Pairagon only, and the curves with triangle dots are of N-SCAN_EST+Pairagon. Solid lines are for sensitivity and dotted lines are for specificity. The left most two triangle dots with no cDNAs on the N-SCAN_EST+Pairagon curves stand for pure N-SCAN result.

Chapter 4

Guide Biology Experiments by Summing Over Consistent Prediction Models

Although the ultimate goal of gene structure prediction is to build a system that is accurate enough such that experimental verification and manual curation is no longer necessary, no existing gene prediction system is close to that yet, especially for vertebrate genomes (Brent and Guigo 2004; Brent 2005). All the gene structures predicted computationally need to be verified experimentally. MGC (Mammalian Gene Collection) is such a project trying to identify and accurately sequence at least one full ORF-containing-cDNA clone for each human, mouse, rat and cow gene (Strausberg, Feingold et al. 2002). In order to reduce the time and cost for identifying full ORF-containing cDNAs clones, ESTs are sequenced from both ends of cDNAs. Because of the poly-A tail in the 3'end, the reverse transcription from RNA to cDNA may not extend to the 5' end. So if a cDNA contains the start of translation, it often implies that this cDNA contains a full-length open reading frame (ORF). Gene prediction method can be used to assign a score to each 5' EST aligned to a genomic sequence. This score can then be used as an index of the probability that a cDNA contains a full-length ORF. CDNAs with higher probability containing a full-length ORF will be sequenced with higher priority. Section 4.1 describes the problem of finding a full-length cDNA with EST alignment in a more abstract way. The summation of consistent paths method is then introduced in section 4.2. Section 4.3 covers the simulation results of the Summation of Consistent Paths Method. Section 4.4 discussed the effect of the method on a real data set.

4.1 The Problem

Some biology experiments are expensive and time-consuming, such as finding full-length cDNA clones (Strausberg, Feingold et al. 2002). 5’ ESTs are generated from cDNA clones to evaluate the probabilities that these cDNA clones contain complete ORFs. By analyzing these ESTs, the corresponding clones with high potential to contain complete ORFs are selected. Only these selected clones are sequenced to high accuracy. Gene prediction method can be used in the selection of full-ORF cDNAs. A more general and abstract form of this problem may be described as following.

Input: A sequence S, a gene structure model M for gene prediction and a set of regions R on S.

Output: The probability that R contains a feature A under the probability model M

In full-length cDNA clone selection case, sequence S is a genomic sequence, model M can be the probabilistic gene structure model TWINSCAN, R region are the EST aligned regions, and feature A is the start of translation.

Determining if an EST aligned region (in R) contains the start of translation (A) or not can be done by running gene prediction on this sequence S, and comparing the prediction results and region in R, to see if this region contains the feature A. But if we already know that regions in R can only be feature A, or B in a model M, then regions in R themselves contain very important information for gene prediction. The method only comparing the gene prediction and regions in R is called a simple method. We can do better if we use the information that R can only be feature A or B from the beginning of the gene prediction. The success of integrating EST alignments with TWINSCAN shows that gene prediction can be improved if EST alignments are already known.

4.2 The Summation of Consistent Paths Method

The algorithm developed here uses the EST alignment and the gene structure model to rank the potential of EST aligned regions containing start of translation. The method is called summation of consistent paths method. A trellis data structure is used for gene prediction. A path in this trellis corresponds to a gene structure prediction (Figure 4-1). In TWINSCAN, only the maximum likelihood path is selected in the prediction. However, alternative splicing is very common in mammalian genomes, about 30% to 60 % of human genes are estimated to have distinct alternative splicing (Modrek, Resch et al. 2001; Modrek and Lee 2002; Lee, Atanelov et al. 2003; Modrek and Lee 2003). In this algorithm, all the consistent paths will be considered. The paths consistent with EST alignments will be classified into two categories: one is for the paths containing the start of translation; the other is for the paths not containing the start of translation. This summation method will compute the summation of probabilities of all the consistent paths containing the start of translation, which is called Yes_probability, and the summation of probabilities of all the consistent paths not containing the start of translation, which is called No_probability. The ratio of these two probabilities summation is used to represent the probability of the EST alignment region containing the start of translation. In practice, all the probabilities are represented by their log values, also called scores. The score for Yes_probability is called *Yes_score*, and the score for No_probability is called *No_score*. The ratio of Yes_probability and No_probability becomes the difference of the *Yes_score* and *No_score*. The advantage of this summation method is that all the suboptimal paths will be well represented in the final score if these suboptimal paths have very close scores to the optimal one. This is very important when an EST alignment is derived from an alternative splicing which is not predicted in the optimal path. Forward algorithm was used to compute the score summations.

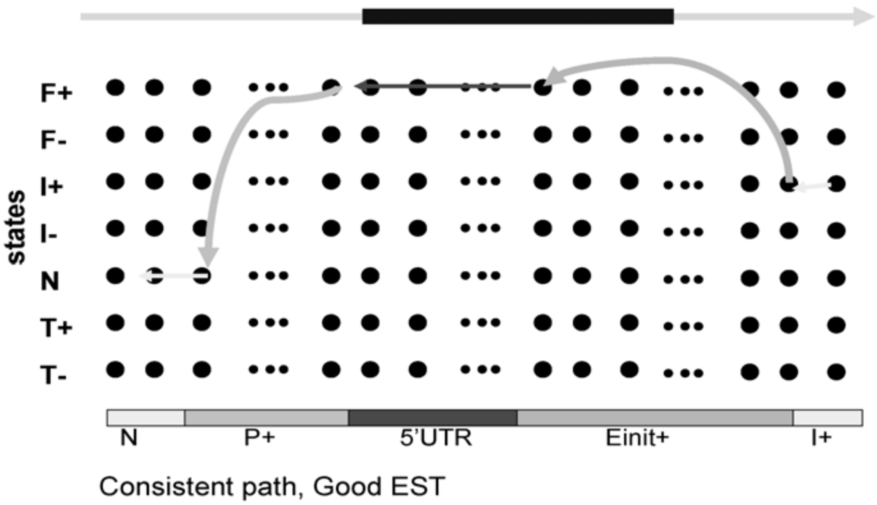


Figure 4-1. An example of consistent path with EST alignment region containing the start of translation.

The upper line stands for an EST (the black box) aligned to a genomics sequence. Only UTR, intron and intergenic states are shown in the trellis. Exon states are represented by arrowed lines from introns to 5' UTRs, introns to introns or 3'UTRs to introns. Each path in the trellis corresponds to a gene structure prediction. If a path predicts the entire EST aligned region either as UTR region or coding region, then this path is a consistent path. The score summations of the two classes of consistent paths are calculated for each trellis cell (big black dot in the figure) by forward algorithm for HMM. We can get the overall summation score for two classes of paths by adding the two scores in the last column of the trellis.

The complexity of algorithm to compute the summation scores is the same as TWINSCAN, which in practice, grows about linearly with the input sequence length. The additional time for EST alignments depends on the program used to do spliced alignment.

4.3 Simulation Result

The efficiency of this algorithm was tested in a simulation data set in which 364 single gene sequences were from GENSCAN training data sets. Two artificial ESTs for each gene were generated: One is “good”, which is from 100bps upstream of the start of translation and with size 500 bps; the other is “bad”, which is from 100 bps downstream of the start of translation and also with size 500 bps. Thus, the “good” EST contained the translation start, and the “bad” EST did not contain the start of translation. The algorithm assigned a score to each EST, and then a threshold could be picked to separate good ESTs (predicted) from bad ESTs (predicted). The efficiency of this algorithm could be tested by the ratio of good ESTs being predicted as good ESTs (true positive) and the ratio of bad ESTs being predicted as good ESTs (false positive).

If value 1.2 is chosen as the threshold for *Yes_score – No_score*, all the ESTs with *Yes_score-No_score* larger or equal to 1.2 are predicted as good EST or otherwise as bad EST. 182 (50%) of the good ESTs are predicted as good ESTs, while 92(25%) of the bad ESTs are predicted as good ESTs. If we have 100 good ESTs and 100 bad EST before the procedure, we need 200 experiments to get 100 full ORF-containing cDNAs. The success rate is about 50%. After the procedure, the number of clones will be sequenced is 50+25 = 75, and 50 will get full ORF-containing cDNA. So the success rate goes up to 67%. If we have 100 good ESTs and 900 bad EST before the procedure, we need 1000 experiments to get 100 full ORF-containing cDNA. The success rate is about 10%. After the procedure, the number of clones will be sequenced is 50+225 = 275, and 50 will get full ORF-containing cDNA. So the success rate goes up to 18%, almost doubled.

4.4 Experiment Result

A real data set was constructed to test the efficiency of the algorithm. 26,991 ESTs whose cDNA clones had been picked to sequence to high accuracy in MGC were downloaded from ftp://ftp1.nci.nih.gov/pub/MGC/ on 10/29/2002, and 13,526 RefSeq genes were downloaded from

http://genome.ucsc.edu/goldenPath/28jun2002/database/refGene.txt.gz. 3611 ESTs had matched region larger than 95% when 26,991 ESTs were aligned to RefSeq cDNAs using BLASTN. These 3611 ESTs were divided into good and bad groups by aligning them using BLAT to the genomic locus of the RefSeq with an additional 1000 bases at each end. There were 2558 good ESTs and 1053 bad ESTs, and each EST had its RefSeq DNA sequence. The summation algorithm has been tested on this real data set. The results are shown in (Table 4-1). 52% of the Good ESTs were predicted as Good ESTs while only 29% of the bad ESTs were predicted as Good ESTs. The results are consistent with our simulation results.

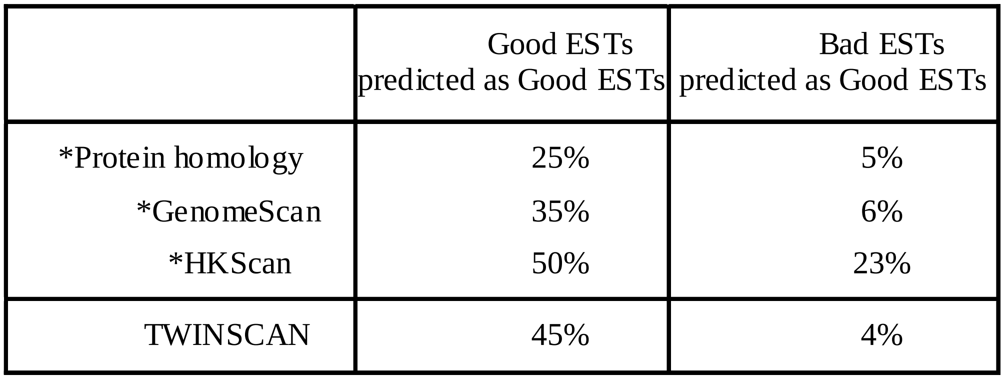
Table 4-1. Efficiency of summation method.
3611 RefSeq genes, 2558 good ESTs and 1053 bad ESTs.

	Good EST	Bad EST
Predicted as Good EST	51.96%	28.84%
Predicted as Bad EST	48.04%	71.16%

The effect of simple method on this data set was examined by running TWINSCAN on these genomic sequences aligned by the RefSeq sequences with 1000 additional bases on both ends and comparing the prediction results and the EST alignments. The result was compared to algorithms used in MGC project. Results are shown in Table 4-2. Since the total probability is the probability of all the gene models that are consistent with the EST alignment, the longer an EST is the higher will be the total score. There is no normalization procedure involved here. It might help if scores is normalized by the length of genomic sequences and ESTs. Different strategy of normalization might also improve the method's accuracy. The gene structure modeling for gene prediction has been improved in many aspects, such as explicit intron length distribution model and non-canonical splicing site models. Eventually, these improvements in gene structure modeling can impact the accuracy of the full ORF-containing cDNA clone selection too.

Table 4-2. Efficiency of simple method.

3611 RefSeq genes, 2558 good ESTs and 1053 bad ESTs were used for the TWINSCAN simple method. * For algorithms used in MGC clone selection algorithms, the dataset contained 3255 RefSeq genes, and each RefSeq gene has one good EST and one bad EST(Strausberg, Feingold et al. 2002).



	Good ESTs predicted as Good ESTs	Bad ESTs predicted as Good ESTs
*Protein homology	25%	5%
*GenomeScan	35%	6%
*HKScan	50%	23%
TWINSCAN	45%	4%

Chapter 5

EST-to-Genome Alignment with Sequencing Quality Values

Accurate alignment of ESTs, cDNAs, mRNAs or protein sequences to genomic sequence is critical to many bioinformatics applications such as gene prediction, single nucleotide polymorphism (SNP) detection and alternative splicing finding. It is also called spliced alignment because when a transcript product is aligned to a genomic sequence, the introns appear as gaps in the alignment. This Chapter starts with a brief review of spliced alignment algorithms. Then sequencing quality value is introduced in Section 5.2. Single nucleotide polymorphism is described briefly in Section 5.3. Section 5.4 describes a graphical model to represent the error patterns in a correct alignment arising from sequencing error and polymorphism (or evolution divergence if aligned cDNA and genomic sequence are from different organisms). A pair Hidden Markov Model (pair HMM) for spliced alignment is covered in Section 5.5. A bootstrap parameter estimation procedure is introduced for QPAIRAGON in Section 5.6. Section 5.7 covers the stepping stone algorithm and intron cutout algorithm to improve the speed and reduce the memory requirement of the pair HMM. The experiment data set for the program is presented in Section 5.8. Section 5.9 introduces a novel accuracy measurement for spliced alignment algorithms based on known single nucleotide polymorphisms. Section 5.10 compares the results of different alignment algorithms on the data set described in Section 5.9. Section 5.11 discusses the effect of alignment programs on gene prediction by replacing the default alignment program BLAT with the new alignment program QPAIRAGON. Section 5.12

compared the effects of different spliced alignment programs on crossing-species alignment. Some QPAIRAGON alignment examples are shown in Section 5.13. The behavior of QPAIRAGON is discussed with these examples. Section 5.14 introduces posterior probability to measure the reliability of alignments. The last section contains conclusions and discussions.

5.1 Spliced Alignment

Aligning a transcript product, such as an EST, cDNA and mRNA to a genomic sequence is different from aligning a genomic sequence to another genomic sequence. When a transcript, an EST for example, is aligned to a genomic sequence, segments of it will be spliced at the splice sites and aligned to different regions of the genomic sequence. If an EST sequence is perfect and there is no SNP difference between the EST and the genome sequence, then its alignment to a genomic sequence will provide unquestionably exact exon-intron structure of the gene regions it covers, including both the coding regions and UTRs. However, EST-to-genome alignment is a difficult problem because of the low sequence quality, SNPs and other issues such as strand uncertainty (Kan, Rouchka et al. 2001) and all kinds of contaminating sequences, such as those cloning vectors, intron sequences, chimeric sequences, unspliced pre-messenger RNAs and genomic DNAs (Wolfsberg and Landsman 1997).

A number of algorithms have been developed for EST-to-genomic alignment (Huang, Adams et al. 1997; Mott 1997; Florea, Hartzell et al. 1998; Usuka, Zhu et al. 2000; Wheelan, Church et al. 2001; Kent 2002; Brendel, Xing et al. 2004; Zhang and Gish 2006). In general, these existing spliced sequence alignment methods can be classified into two categories. One uses heuristic strategy like BLAST (Altschul, Gish et al. 1990; Altschul, Madden et al. 1997) and the other uses dynamic programming to find an optimal solution. Sim4(Florea, Hartzell et al. 1998), Spidey (Wheelan, Church et al. 2001) and BLAT (Kent 2002) belong to the first category. Sim4 finds identical seed words of length 12, and then extends on both directions with a greedy algorithm. Shorter seeds are used to fill unaligned regions. BLAT, which is short for "BLAST-like alignment tool", was designed to align millions ESTs against the whole human genome. It speeded up the alignment by indexing and storing the whole human genome in memory. EST_GENOME (Mott 1997), dds/gap2 (Huang, Adams et al. 1997), GeneSeqer (Usuka, Zhu et al. 2000) and EXALIN (Zhang and Gish 2006) belong to the second category. EST_GENOME uses dynamic programming to find the optimal spliced alignment for each given score scheme. It also incorporates splice site signals by preferring GT at donor and AG at acceptor site. All other methods in the second category use a 2-stage strategy to speed up the alignment, though dynamic programming is used in the second stages to improve the alignment quality. EXALIN incorporates PSSM splice site models into the dynamic programming it employ for spliced sequence alignment. GeneSeqer uses a Hidden Markov Model on the second stage of fine tuning. The heuristic methods are generally much faster than the full dynamic programming methods, though their alignment accuracy is lower. The alignment accuracy can be improved by integrating splicing site models (Usuka, Zhu et al. 2000; Zhang and Gish 2006).

High quality EST-to-genome alignment is critical for many computational biology applications. Any increase in the alignment accuracy will improve our understanding of genomic biology procedures such as splicing and consequent wet lab experiment results as well. However, a correct alignment of an EST base to a genome can be a non-match. The two major reasons are the sequencing errors and SNPs (see Section 5.3). The sequencing quality values introduced in the next section can be an extra resource to improve the alignment quality.

5.2 Sequencing Quality Values

Sequencing quality values are computed by base-calling programs such as PHRED (Ewing and Green 1998; Ewing, Hillier et al. 1998). A quality value

$$q = -10 \times \log_{10}(p) \qquad (5.1)$$

is assigned to each base call, where p is the estimated error probability for the base call. The quality values give a measure of the reliability of the sequence and they can greatly improve the assemble accuracy in some sequencing projects. For EST-to-genome alignment, if the quality value at each base is used, then the alignment will combine not only the alignment of the EST sequence and the genome sequence, but also the reliability of the EST sequence. A match in a high-quality region can have a higher score (probability) than a match in a low-quality region; and a mismatch, insertion or deletion in a high-quality region can have a higher penalty than a match in a low-quality region. In this way, we may expect that the effect of the sequencing error on the EST-to-genome alignment can be reduced. However, none of the existing spliced alignment algorithms has used the EST sequencing quality values.

NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces) is a repository of the raw sequence traces. It contains both sequence and the quality value of the EST reads. Up to September 25, 2005, 1,953,938 human EST reads could be downloaded, of which 492,492 reads had non-zero quality values and their quality value sequence lengths were consistent with length of the read sequences. Figure 5-1 shows the average quality value distribution counted from these reads. For each EST, only the largest region with less than 25 expected error bases is counted. The average quality values from the first 150 bases and last 400 bases of this region are displayed on left and right side of the figure respectively. If a region was shorter than 150 bases, then all its bases were counted in the first 150 base. Similarly, all the bases were counted in the last 400-base statistics if a read is shorter than 400 bases. The figure shows the general quality value distribution in a read sequence. A read sequence has lower quality values at both ends, and the quality values become higher as it gets away from both ends.

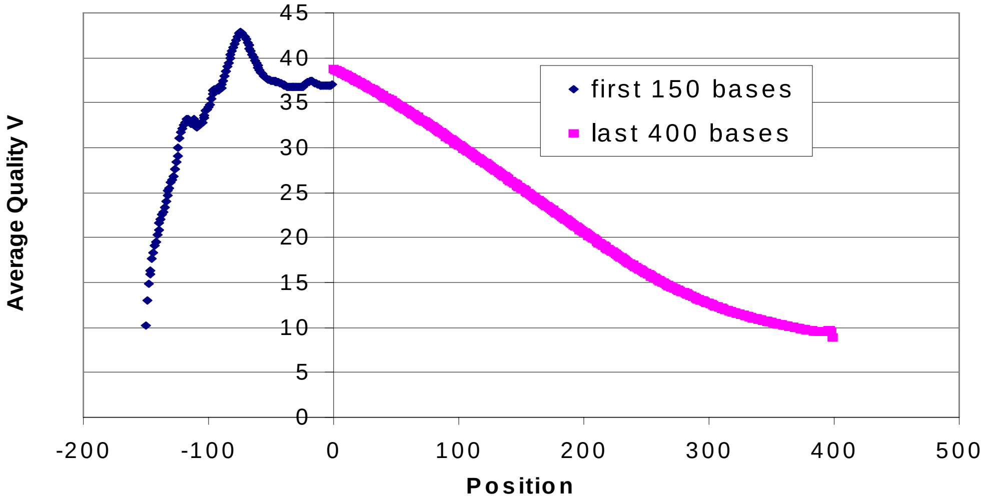


Figure 5-1. Sequencing quality value distribution.

The average quality value at each position is shown in this picture. The left part (-150 to 0) represents quality value distribution on the first 150 bases. The right part (0 to 400) represents quality values distribution on the last 400 bases. The average values were counted from the human reads with non-zero quality values.

5.2.1 Sequencing Error Patterns

Different types of errors may happen during sequencing. Error types include substitutions, insertions, deletions and ambiguous bases. Although the error rate generally decreases when the sequencing quality value grows, it is instructive to investigate the distribution of different error types. Ewing and Green reported the sequencing error type distributions of automated sequencer traces using Phred on some mammalian clones and *C. elegans* clones (Ewing and Green 1998). Based on the number of errors reported in the paper, the overall error distribution can be derived for different patterns. 71% of the errors are substitutions, 17% are deletions, 8% are insertions and 4% are undecided bases. These numbers of error pattern distribution might vary due to many elements in different sequencing projects. Since these were the only data reported at the time, they were used to guide parameter estimation for the new spliced-sequence alignment method, which is covered in Section 5.6 below.

5.3 Single Nucleotide Polymorphism (SNP)

A Single Nucleotide Polymorphism, or SNP, is a small genetic change or variation that occurs within a person's DNA sequence when a single nucleotide replaces one of the other three nucleotides. Generally, a SNP only presents in a small percentage of the whole population. Since the protein coding regions are only about one percent of the whole human genome, most of the SNPs occur in non-coding region. Many common human diseases are believed to be caused by some SNPs especially those in coding regions. Therefore, SNPs are very important for disease diagnoses and drug development. Millions of SNPs in the human genome have been found (Sachidanandam, Weissman et al. 2001). A SNP can be represented by letters separated by a "/", such as "A/G" stands for an "A" being replaced by a "G" or a "G" by an "A". Among all the known SNPs, the most common SNP is "C/T".

ESTs are deposited into database from different research groups on different projects all over the world, therefore they may come from a different individual than the genome does. When an EST is aligned to the genome, and they are from different individuals, a correct alignment of an EST base to a genome can be a non-match even if the EST itself is error free, since it may be caused by a SNP between the EST and the genome.

5.4 A Graphical Model for Error Patterns in Correct Alignments

In a HMM, each state can have a state-specific model for emission. A graphical model can be introduced for each state in the pair HMM to integrate the sequencing quality values into sequence alignment. Figure 5-2 shows the model and null-model to represent the error patterns of correct alignments. RG stands for the true sequence of the reference genome, from which the genomic sequence is derived, EG for the true sequence of the EST genome, from which the EST sequence is derived, and EC for the EST base call obtained by sequencing procedure. RG and EC are observable while EG is a hidden variable. EC is determined by the EG and the quality value state “qual”. When aligning a base of an EST sequence (ECs from an EG) to a base of a genomic sequence (from an RG), a correct alignment can be other than a match. If the EG and RG are from the same species, without considering contaminations, a correct non-match alignment may be caused by sequencing errors (from EG to EC) and/or SNPs (between RG and EG) or other elements such as RNA editing. But sequencing error and/or SNPs account for the vast majority of the differences. If the EG and RG are from different species, it may be caused by sequencing error from EG to EC and/or evolutionary divergence between the EG and RG. The undirected edge between RG and EG represents an association relationship between them. The graphical model shown on the left side of the figure represents when the reference genome and EST genome are not independent to each other. The null-model shown in the right side of the figure represents that the reference genome and EST genome are independent to each other, which means two random sequences are being aligned. As in many HMMs, the optimal likelihood ratio of the model and null-model may be computed to represent the best alignment of the two sequences.

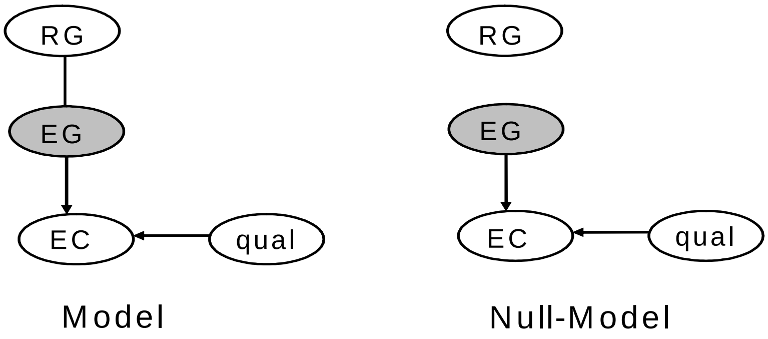


Figure 5-2. A Graphical model for EST-to-genome alignment with quality value sequence. “RG” stands for the reference genome, “EG” for the EST genome, “EC” for EST base calls, and “qual” for quality values of the base calls. Left part of the figure is the model and the right part is the null-model. The undirected edge in model means that “RG” and “EG” are not independent of each other. In the null-model, the “RG” and “EG” are independent to each other.

Under this graphical model, the reference genome base (RG) is independent of the EST base call (EC) given the EST genome base (EG). Therefore, the probability of observing an EC base and an RG base given a quality value *q* can be expressed as

$$\Pr_{\text{mod}}(EC, RG \mid q) = \sum_{EG} \Pr(EC, EG, RG \mid q) = \sum_{EG} (\Pr(EC, EG \mid q) \Pr(RG \mid EG)) \quad (5.2);$$

Under the null model, because the reference genome base (RG) is independent of the EST genome base (EG), the probability of observing an EC and RG given a quality value q is

$$\Pr_{null\_\text{mod}}(EC, RG \mid q) = \sum_{EG} \Pr(EC, EG, RG \mid q) = \sum_{EG} (\Pr(EC, EG \mid q) \Pr(RG)) \quad (5.3).$$

In practice, log-scale scores replace the probabilities. The score of the model and null-model could be represented as

$$S(EC, RG \mid q) = \log \frac{\sum_{EG} (\Pr(EC, EG \mid q) \Pr(RG \mid EG))}{\sum_{EG} (\Pr(EC, EG \mid q) \Pr(RG))} \qquad (5.4).$$

When aligning human ESTs to human genomic sequences, $\Pr(EC, EG \mid q)$ can be calculated from sequencing error pattern distribution; $\Pr(RG \mid EG)$ can be calculated from human SNP data and Pr(RG) can be estimated from the whole human genome sequence.

From equation (5.1), the overall probability that a nucleotide base with quality value *q* is an error is $p = 10^{-q/10}$. The overall error includes 12 different types of substitutions, 4 types of insertions and 4 types of deletions, and ambiguous EST bases. The following equation (5.5) describes how to compute $\Pr(EC, EG \mid q)$ from the sequencing error pattern distribution reported in Section 5.2.1.

$$\Pr(EC,EG\mid q)=\begin{cases} \frac{1}{4}(1-10^{-q/10}) & if \quad EC=EG \\ \begin{cases} \frac{1}{12}\,substitution & if \quad EC\neq"\_",EG\neq"\_" \\ \frac{1}{4}\,insertion & if \quad EC\neq"\_",EG="\_" \\ \frac{1}{4}\,deletion & if \quad EC="\_",EG\neq"\_" \end{cases} \end{cases} \qquad (5.5).$$

The parameter estimation method introduced above assumes linear gap penalty, i.e. there is no differentiation between gap open penalty and gap continuous penalty. The parameter estimated by this method can be used to parameterize a HMM which has equivalent linear gap penalty. Figure 5-3 shows the theoretical alignment scores estimated as mentioned above using the sequencing error distribution from the Phred paper (Ewing and Green 1998) and SNP statistics from the 4.2 million SNPs downloaded from ftp://ftp.ncbi.nih.gov/snp/human/rs_fasta (November 3rd, 2004). When the quality value is getting lower, the gain of a mach goes down and the penalty for a mismatch and indel also becomes smaller. For all quality values, the penalty of the substitution is smaller than the penalty of a deletion and the penalty of a deletion is smaller than that of the insertion. Therefore the substitution is more likely than deletion and deletion is more likely than an insertion for all the quality values. This is consistent with the sequencing error distribution reported in (Ewing and Green 1998).

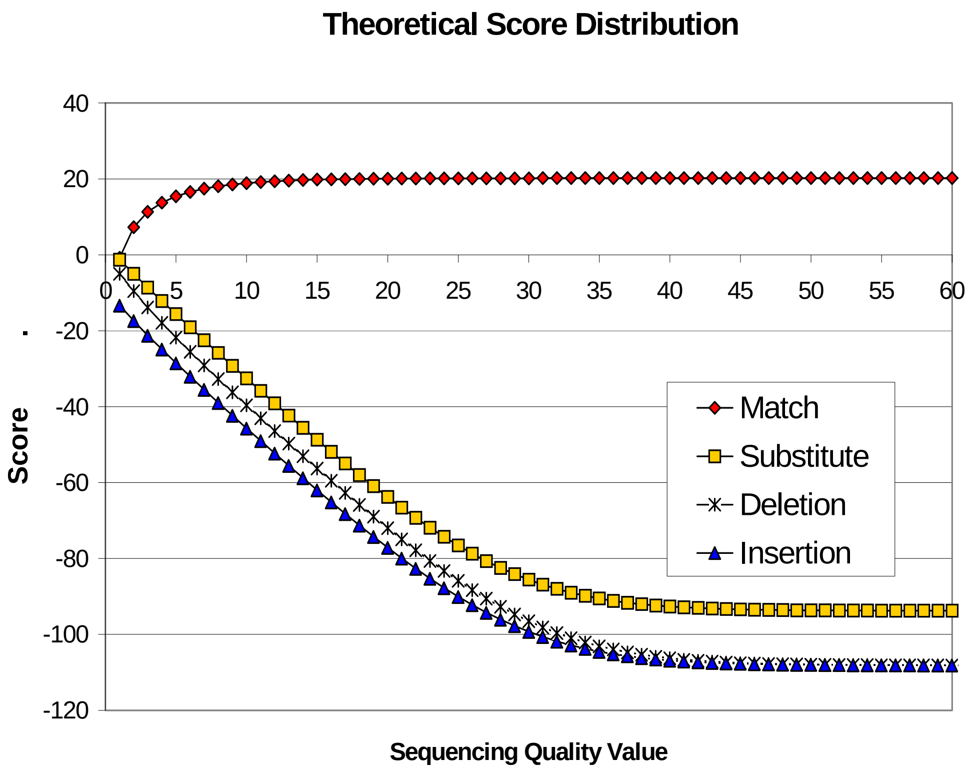


Figure 5-3. Alignment scores estimated from the Graphical model.

Scores are estimated as described in section 5.5. From top to bottom, the line with diamonds is for the matches, squares for mismatches, triangles for cDNA insertions and * for deletions. When quality value is getting lower, the gain of a mach goes down and the penalty for a mismatch and indel also become smaller. All insertion and deletion scores shown here are for length one.

The theoretical parameter estimation approach introduced in this Section does not depend on any existing spliced alignment programs and it has the potential to be extended to cross-species alignment by replacing the polymorphism part with cross-species evolution divergence. However, the score matrix derived by this method only has linear penalty for gaps and the conversion from the scores derived from this graphical model to a pairHMM framework is not unique. The pairHMM framework naturally supports affine score scheme for all states, which differentiates the state opening and state continuation. The parameters for QPAIRAGON can be estimated by this way or, alternatively, by running other spliced alignment program, such as EST_GENOME, on some data set and re-estimating QPAIRAGON parameters from the alignments.

5.5 Pair HMM for Sequence Alignment

Our new EST-genome alignment algorithm uses EST sequencing quality values as well as the EST sequence and the genomic sequences to improve the alignment accuracy. In order to speed up the computing, a two-stage alignment strategy is used by most fast alignment algorithms (Florea, Hartzell et al. 1998; Usuka, Zhu et al. 2000; Kent 2002). The first stage is a fast search stage to find regions of the two sequences that are likely to be homologous. Then, a dynamic programming stage is used to find more accurate and reliable alignments. The system developed here uses existing fast alignment program BLAST to do the first stage fast search. The focus of this section is on the second alignment stage.

5.5.1 Pair HMM for Sequence Alignment

The model used for the second alignment stage was a pair HMM (Durbin 1998). A standard pair HMM for genome-to-genome sequence alignment is reviewed in Section 2.6. For EST-to-genome alignment, the standard pair HMM was modified to include intron states for genomic sequence only. In a pair HMM using sequencing-quality value, the inputs are three sequences: a genomic sequence, an EST sequence, and a quality value sequence for the EST sequence. The diagram of the algorithm is shown in Figure 5-4. “M” is for match and mismatch and “In” for intron state emitting only genomic sequence. “Enter” and “Exit” corresponds to the splice donor site and acceptor site. In practice, they contain the first two bases of 5’ end of an intron and the last two bases at the 3’ end of an intron. “G” and “C” states are for insertions in genomic sequence and EST sequence respectively. “RC” and “RG” states are for random cDNA and random genomic sequences. They are introduced to deal with genomic or cDNA overhang on one or two ends. The little “q”s in the diagram stand for the emission scores of the states. In the states, q_{x_i} is the score of emitting an x_i from the genomic sequence; $q_{y_j z_j}$ is the score of emitting a y_j from the cDNA/EST sequence; and $q_{x_i y_j z_j}$ is the score of emitting a (x_i, y_j) pair from both sequences with z_j as the sequencing quality value at position *j* of the cDNA/EST sequence.

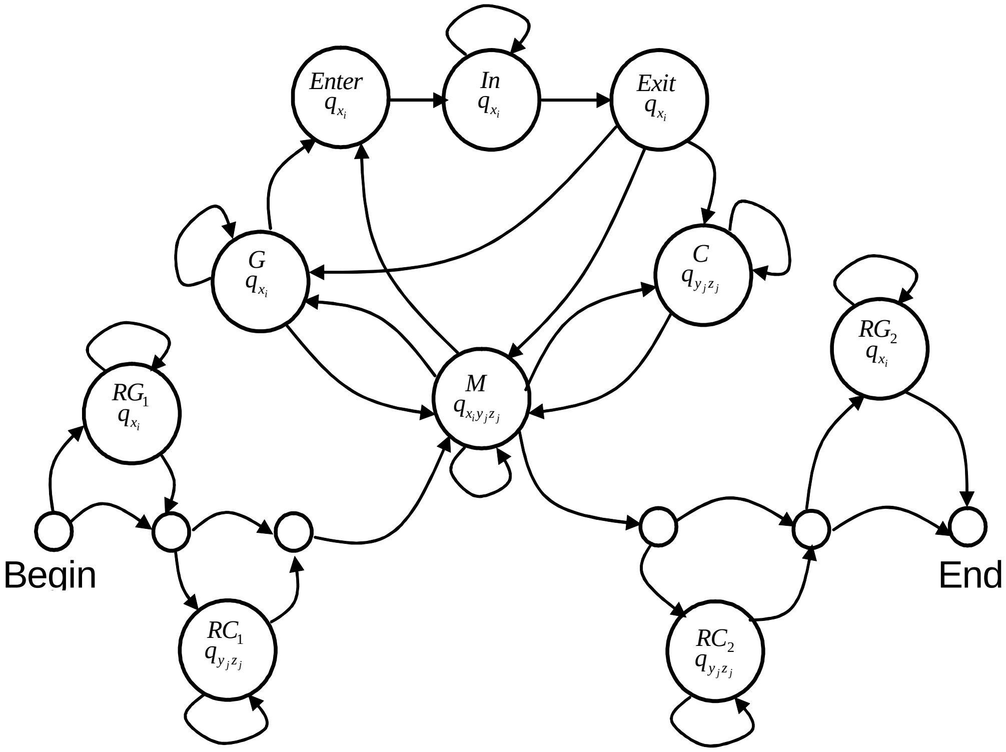


Figure 5-4. Diagram of pair wise alignment HMM for integrating sequencing quality value to spliced alignment.

The small circle states, which do not generate any symbol, are presented for convenience. x_i is the i^{th} base in the genomics sequence, and y_j and z_j are the base and the quality value for the base in the EST sequence. RG1, RC1, RG2 and RC2 are random model states. The alignment result is a local alignment by using these four random model states. M is for match and mismatch; In for intron state in the genomic sequence; and G and C states for inserts in genomic sequence and EST sequence respectively. Note that only genomic sequence has intron state. And the splice site models are enforced in the transition from M state to In state. There are different Intron states representing different type of Intron patterns, and for different orientations. To make it simple, only one Intron state is shown in this figure.

5.5.2 Viterbi Algorithm for Pair HMM

Chapter 2 introduced a pair HMM for sequence alignment. Assuming there is no sequence quality values involved, the most probable path or the best alignment can be computed by Viterbi algorithm for the model shown in Section 5.5.1. If $v^k(i, j)$ stands for the probability of the most likely path that ends in state k having emitted $x_1,..., x_i$ of the genomic sequence and $y_1,..., y_j$ of the EST sequence. Let *n* and *m* be the length of the genomic sequence and the EST sequence respectively, then Viterbi algorithm for the pair HMM in Section 5.5.1 can be expressed as below.

Viterbi algorithm for the pair HMM in Section 5.5.1

Initialization: $v^M(0,0) = Init(M)$, $v^{RG_1}(0,0) = Init(RG_1)$, $v^{CG_1}(0,0) = Init(CG_1)$ and all other $v^k(i,0)$ and $v^k(0, j)$ are set to 0, where $1 \le i \le n$ and $1 \le j \le m$.

Recursion: *i*=1, …, *n*, *j*=1, …, *m*;

$$v^k(i,j) = \begin{cases} \max_l v^l(i-1, j-1) * a_{lk} * e_k(x_i, y_j), & if\ k = M \\ \max_l v^l(i-1, j) * a_{lk} * e_k(x_i), & if\ k = G, In, RG_1, RG_2, Enter \quad or \quad Exit \\ \max_l v^l(i, j-1) * a_{lk} * e_k(y_j), & if\ k = C, CG_1 or \quad CG_2 \end{cases}$$

Termination: $v^{end} = \max_l v^l(n, m)$.

a_{lk} is the transition probability between state *l* and *k*. $e_k(x_i, y_j)$ is the probability of state k emitting (x_i, y_j). $e_k(x_i)$ is the probability of state k emitting x_i only and $e_k(y_j)$ is the probability of state k emitting y_j only. $\max_l$ is for all legal transitions to k (from l).

The integrating of quality values into the HMM can be considered as follows. For a state emitting a symbol y_j or a pair of symbols (x_i, y_j), its emission probability will be changed to a quality-dependent probability, i.e. $\Pr(y_j \mid z_j)$ replaces $\Pr(y_j)$ and

$\Pr(x_i, y_j \mid z_j)$ replaces $\Pr(x_i, y_j)$. If a state emits an EST/cDNA base, the transition probability out of this state is also dependent on the neighboring quality value. Thus, in Viterbi Algorithm for the pairHMM, the recursion for $v^k(i, j)$ will be the same except that the state transition probability and emissions probability are changed to quality dependent.

The splice site models are enforced in the transition between the “M“ state and “In” state. Canonical intron structures like GT/AG and GC/AG can have much lower intron penalties than non-canonical intron structures like “AT/AG”.

5.5.3 Algorithm Complexity

Since a spliced alignment algorithm needs to deal with thousands or even millions of EST sequences, speed is an important issue. The computation complexity of a naive implementation is $O(N^2GE)$, where N is the number of states in the HMM, and G and E are the length of genomic sequence and EST sequence respectively. The memory requirement is $O(NGE)$. For large size sequences, this may be a problem. For example, if G=100kb, E=500b, N=15, and each cell in the trellis requires 12 bytes, then 9 GB memory is needed to run this alignment for a naive implementation. It is not rare that the value of G is in several hundred Kb. Section 5.5 presents a stepping stone algorithm to constrain the alignment within a bounded area so that to speed up the program and reduce memory requirement as well. Section 5.6 introduces a method to cutout long intron regions under a certain condition which may further increase the speed and reduce the memory requirement.

5.6 Bootstrapping Parameter Estimation for QPAIRAGON

Since no existing EST-to-genome alignment program has used the sequencing quality value, there was no reliable training data for QPAIRAGON. A bootstrapping parameter estimation is introduced in this section. EST_GENOME alignments on the Set B described in Section 5.9 were used to re-estimate parameters for QPAIRAGON.

A simple Gaussian smoothing was applied to smooth the counts with different quality values. For each state in the pairHMM, a Gaussian window of size 9 with standard deviation 2 was slid through counts for all the quality values. In other words, assume m_q is the original number of matches with quality value q and $g(i) = \frac{1}{A}\exp^{-\frac{i^2}{2*4}}$, where $A = \sum_{i=-4}^{4}\exp^{-\frac{i^2}{2*4}}$ and $-4 \le i \le 4$. The count for quality value q after smoothing becomes $\tilde{m}(q) = \sum_{i=-4}^{4} m(q+i) * g(i)$. We assume m(q)=m(0) for $q < 0$, and $m(q) = m(q_{\max})$ for $q > q_{\max}$.

Figure 5-5 shows the re-estimated alignment scores from EST_GENOME alignments on three-fourths of Set B. The scores shown in the figure are for match, mismatch, insertion open, deletion open, insertion continuation and deletion continuation. When the quality value decreases, the match score decreases and mismatch penalty decreases. It also shows that short indels are preferred when quality values decrease since the indel continuation penalties increase. Compared to the theoretical scores shown in Figure 5-3, the re-estimated penalties for mismatches and indel opens are bigger than their theoretical values especially on low sequencing quality region. This may be caused by not including many low quality value regions in the EST_GENOME alignment results. The lines for indel open and indel continue show that when the quality value decrease, the probability of short indel increase since the indel continue penalty grows while the indel open penalty decreases.

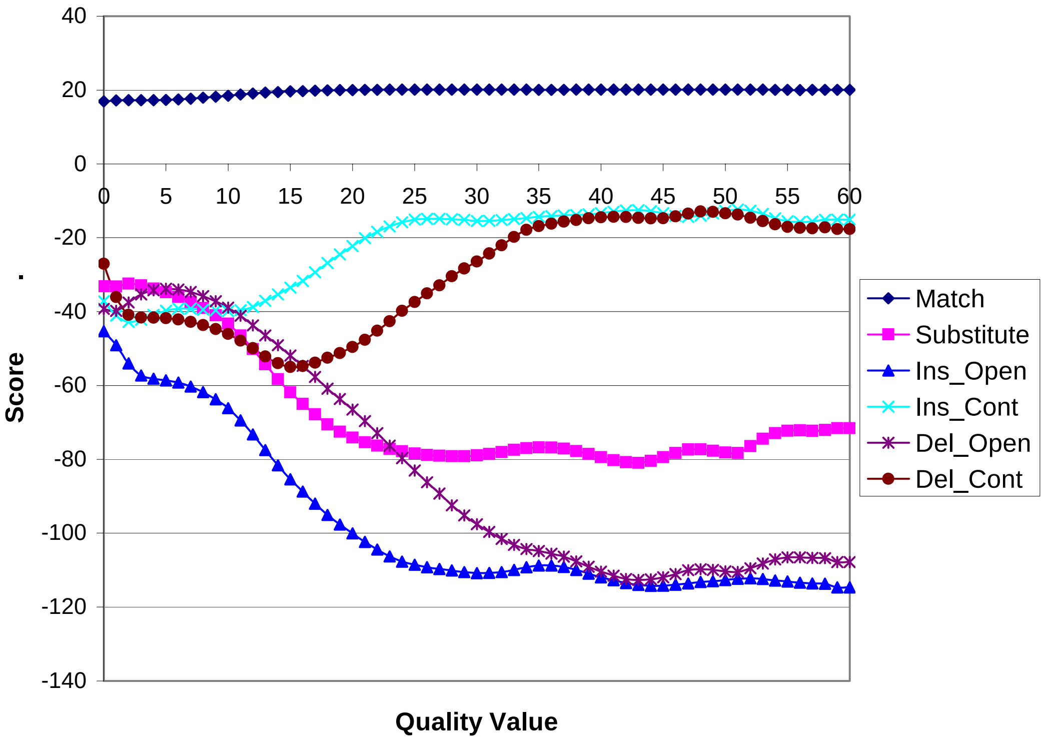


Figure 5-5. Re-estimated alignment scores from EST_GENOME alignments. The training set was three fourth of the EST_GENOME alignments on data set B.

5.7 Stepping Stone Algorithm and Cutout Intron Algorithm

A naive implementation of a pairHMM requires $O(N^2GE)$ and $O(NGE)$ for time and space complexity. The stepping stone algorithm and cutout-intron algorithm are the algorithms introduced to reduce both the computing time and memory requirements.

5.7.1 Stepping Stone Algorithm

The stepping stone algorithm was a heuristic algorithm introduced for a DNA-to-DNA sequence alignment pair HMM (Meyer and Durbin 2002). Both its speed and memory will be scaled linearly with the sequence length for a DNA-to-DNA sequence alignment if the two DNA sequences are homologous. The stepping stone algorithm was modified for EST-to-genome alignment with quality values. The main idea of a stepping stone is that strong similarity between subsequences of a cDNA and a genomic sequence can be used as guidelines to restrict the space of Viterbi algorithm. A simple example of a stepping stone run with two subsequences of strong similarity is shown in Figure 5-6. Highest scoring pairs (HSPs) are local alignments with score higher than a certain threshold. They are diagonal lines in the (X, Y) plane. In the stepping stone algorithm, the pins (the black dots in figure) need to be generated first. A pin is a point in a HSP with at least a certain length, for example 20 bases, of high quality alignments on both directions. Pins can be determined by a fast sequence alignment program first. They stand for regions with highly reliable alignments. In our implementation, HSPs were generated first by running BLASTN for the cDNA and genomic sequences. The subsequence with the strongest similarity (the HSP with highest score from BLASTN) was picked first, and then the best HSP consistent with the picked HSPs was picked from the remaining HSPs. Here, “consistent” means that the two HSPs do not overlap with each other more than a certain number of length, for example 10 bases, in either genomic or cDNA sequence coordinates. Each HSP is checked to see if there is a high quality alignment region (with at most two non-matches in a continuous 40-base region) starting from each end of the HSP. If a high quality alignment region is found, the middle point of a high quality alignment region is called a pin. Rectangles were then created based on the pins, with additional 20 bases on each side of a pin. The additional 20 bases are added so that the ends of exons can move freely across the rectangle to the adjacent exon. Viterbi algorithm ran inside those small rectangles. When computing the optimal path in the next small rectangle, the values computed in the small overlapped square between the two rectangles will be used to initialize the scores for the next rectangle. This procedure is repeated until the ends of the cDNA sequence and the genomic sequence are both reached. The optimal path is then retrieved by tracing back from the bottom-right corner to the top-left corner.

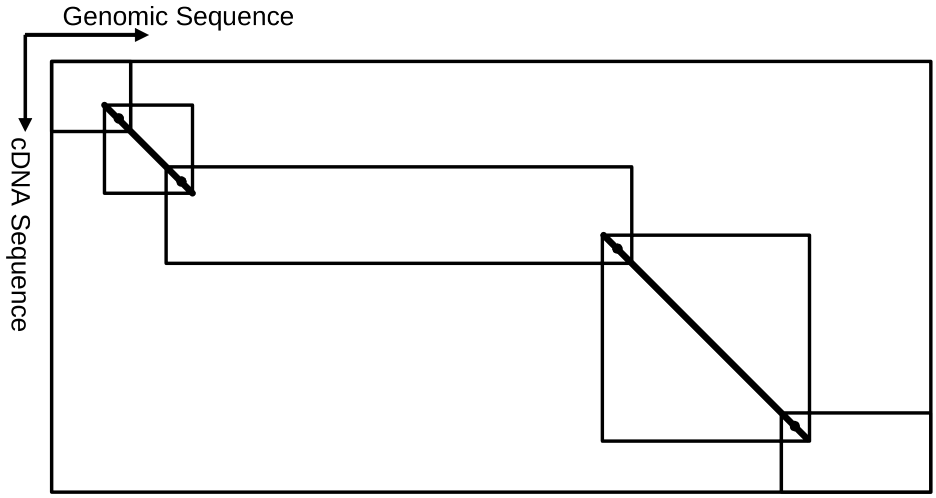


Figure 5-6. An example of Stepping Stone Algorithm with two regions of strong similarity. This Figure shows constraints on the 2-dimension plane spanned by a genomic sequence X and a cDNA sequence Y. The third dimension is the state dimension, on which there is no restriction. The thick diagonal lines strand those strong similarity subsequences determined by the best HSPs in BLASTN. The black dots near the end of the diagonal lines are the alignment pins. The actual search only uses these regions inside the rectangles.

The benefit of the stepping stone algorithm is that the speed will be increased and space requirement will be reduced significantly. More specifically, the time for an alignment of a cDNA and a genomic sequence depends on the number and position of the pins. The memory requirement is reduced by about half with one pin near the middle area, and about two third if two pins appear evenly along the diagonal line. The speed is brought up approximately two and three times respectively as well. In general, if there are *n* pins distributed roughly even in the 2-dimension plane, both the space and time requirement will be reduced to *1/n* of the original ones. In human ESTs, the average number of HSPs per ESTs is about 3. Each HSP generally produces one to two pins, thus we expect to have 3 to 6 pins for each alignment with stepping stone algorithm. Therefore, the total saving on time or space is about 3 to 6 times.

5.7.2 Cutout Intron Algorithm

The speed and memory requirement can be further improved by cutting out large intron regions under certain circumstances. Figure 5-7 shows an example of how an intron region can be cut out. There are two high quality aligned regions (HSPs A and B in the figure) with overlapped ends in cDNA coordinates. If in both A and B, the regions near the overlapped ends are all matches, which means both of them are high quality alignments between the genomic sequence and the cDNA sequence, then a potential exon in the region between the two HSPs can only be generated by moving matches from one or two of the neighboring HSPs. The penalty in overall score by introducing an extra intron is $P = T_{M,In} + T_{In,M}$, where $T_{M,In}$ is the transition score from a match state to an intron state and $T_{In,M}$ for a transition from an intron state to a match state. The gain in overall score by introducing an extra intron is the summation of compensation for the intron donor and acceptor splice sites which are noted as D and A. In the model shown in Section 5.5, the intron and match length distributions are all geometric. The values of D and A for a canonical intron are about 40 maximally, while the penalty P is about 200 if the average intron length and exon length are 5400 and 230 respectively. Overall gain in score by introducing an extra intron is about -120. Since the final alignment is the optimal path within the constrained area, extra intron is not preferred for this region. Therefore, the region between these two high quality HSPs can only contain an intron instead of two or more. We may cut out this region safely if the genomic coordinates of HSPs A and B are sufficiently different from each other. The program has an option to activate this operation. If this option is activated, all input HSPs are checked for their quality and relative positions to each other. If two input HSPs with high alignment quality are overlapping to each other in cDNA coordinates while the difference between their genomic coordinates is larger than 40 bases, the region except extra 20 bases on both ends can be cutout. The score of cutout intron region is compensated in the intron state of the first position flanking the cutout region. The cutout intron regions will be recovered before the final alignment result is reported.

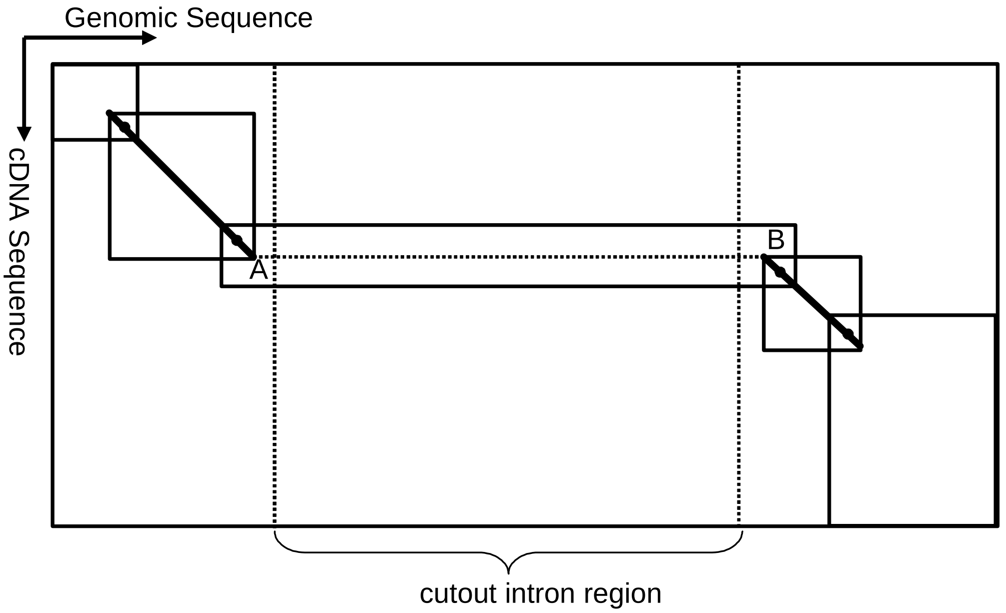


Figure 5-7. Intron Cutout Algorithm.

5.7.3 The Improvement from Stepping Stone Algorithm

A program called QPAIRAGON has been developed based on the algorithms described as above. It may be run with or without the option for stepping stone algorithm. The computing time was collected for QPAIRAGON on Set B with or without the stepping stone algorithm option. The average time for an EST-to-genome alignment in Set B is 256 second without stepping stone algorithm and 37 seconds with the stepping stone algorithm. With 2GB memory limit, 3152 of the total 3214 ESTs are aligned with stepping stone algorithm while only 1334 can be aligned without using stepping stone algorithm.

5.8 Experiment Data Sets

The 492,492 human EST reads with non-zero quality values were aligned to human genome sequences by BLAT. The alignments were then clustered based on their genomic coordinates. An EST alignment was added to a cluster if its genomic coordinates overlapped with at least one base. Then the start and end positions of the cluster were determined by the smallest and largest coordinates of all the ESTs included in the cluster. At the end, there were 4202, 2302, 1032, 553 and 868 clusters in human genome chromosome 1, 5, 20, 21 and 22 respectively. No cluster overlaps any other cluster. Two datasets were created from these EST clusters. Set A contained all the ESTs and their aligned genomic sequences which included 10,000 bases on both ends in addition to the aligned regions. For Set B, the best EST alignments with largest match size inside each cluster was picked first, then the next best one that did not overlap with the picked ones was selected. This procedure was repeated until all the EST reads in the cluster were checked once. Set B contained selected ESTs only and their genomic sequences. The characters of each data set are shown in Table 5-1, Table 5-2 respectively. The genomic sequence parts of these two sets were the same. QPAIRAGON parameters were estimated from Set B and tested on Set B in a 4-fold cross validation. Set A was used to test the effect of different splice alignment programs on gene prediction.

Table 5-1. Characters of ESTs in Set A.

This data set contains all ESTs aligned to chromosome 20, 21 and 22. An EST is called aligned if the number of matches is at least 50% of the length of this EST.

	EST Clusters	EST Number	Total EST Bases	Average EST Length	Total Genomic Bases	Average Genomic Length
Chr20	1,032	10,350	9,617,593	929	38,595,912	37,399
Chr21	553	4,188	3,852,454	919	18,806,503	34,008
Chr22	868	9,215	8,470,883	919	29,946,297	34,500
Total	2,453	23,753	21,940,930	924	87,348,712	35,609

Table 5-2. Characters of ESTs in Set B.

This data set contains the best non-overlapping ESTs. The genomic sequences are the same as those in Set A. ESTs and their genomic sequences on chromosome 1 and 5 are added in a similar way.

	EST Clusters	EST Number	Total EST Bases	Average EST Length	Total Genomic Bases	Average Genomic length
Chr20	1,032	1,291	1,222,541	946	38,595,912	37,399
Chr21	553	646	597,563	925	18,806,503	34,008
Chr22	868	1,116	1,063,444	944	29,946,297	34,500
Chr1	4,202	5,124	4,860,839	948	138,740,552	33,017
Chr5	2,302	2,762	2,609,351	944	83,450,782	36,251
Total	8,957	10,939	10,353,738	947	309,540,046	34,558

5.9 Measuring Spliced Alignment Algorithm Accuracy

Many accuracy measurements have been introduced to evaluate spliced alignment algorithms. The most popular one is aligning artificially mutated cDNA sequences with different levels of error to genomic sequences, and counting the number of accurate intron-exon structures based on manually curated intron-exon structure annotations (Florea, Hartzell et al. 1998; Wheelan, Church et al. 2001; Schlueter, Dong et al. 2003). Here, we present a new accuracy measurement, in which the number of matches, mismatches and mismatches explained by SNPs are counted. A mismatch is explained if it happens in a location of a SNP and can be confirmed by the SNP. The advantage of this measure is that it can explain some mismatches that otherwise can not be explained by other accuracy measures and it assesses the alignment quality directly.

5.9.1 The Same Alignment in Different Forms

Some alignments are essentially the same, but different programs create them in different forms. Figure 5-8 shows an example that the EST gaps in an alignment can move around within a certain range, which can eventually affect the mismatch position. However, all the alignments have the same alignment score if the score matrix depends on the genomic sequence and the EST sequence only.

```
    chr22      14418   GGGGTTGGGGGCGGGGGGGGGGGGGGGTTGGTGTTGAG    14455
                       |||||||||||||||| ||||||||  |||||||||||
 Ti|159693084    532   GGGGTTGGGGGCGGGGTGGGGGGGG--TTGGTGTTGAG     567

    chr22      14418   GGGGTTGGGGGCGGGGGGGGGGGGGGGTTGGTGTTGAG    14455
                       ||||||||||||   |||| ||||||||||||||||||
 Ti|159693084    532   GGGGTTGGGGGC--GGGGTGGGGGGGGTTGGTGTTGAG     567

    chr22      14418   GGGGTTGGGGGCGGGGGGGGGGGGGGGTTGGTGTTGAG    14455
                       |||||||||||||||| ||   ||||||||||||||||||
 Ti|159693084    532   GGGGTTGGGGGCGGGGTGG--GGGGGGGTTGGTGTTGAG     567
```

Figure 5-8. Alignment Result Normalization.

Three alignments with the same alignment score shown here have mismatches in different positions. In all three alignments, the first line is a genomic sequence from human genome chromosome 22, the third line is the EST read sequence with id 159693084, and the second line is the alignment sequence. In the alignment sequences, "|" stands for a match and a blank space stands for a mismatch or an indel. "-" in the EST sequence is for a gap. The mismatch at position (14434, 548) can be (14436, 550) as well if the two EST gaps at (14442, 557) and (14443, 558) were moved to position(s) before (14434, 548). Different programs may give different forms to the alignments, but there is no basis for preferring one over another using the information shown here.

Since the SNP location itself depends on alignment similar to the alignment program used here, in these cases, we do not know where the SNP is. Before the alignment results of different programs were compared, they had to be normalized first so that all the movable gaps were in the left most positions. Then the number of explained mismatches was counted for alignments by different programs.

5.10 Alignment Result Comparisons

Sim4, EST_GENOME, PAIRAGON and QPAIRAGON have been tested on the non-overlap set B described in Section 5.9. The parameters for EST_GENOME were “-align true -match 5 -mismatch 11 -gap_panalty 11 -splice_penalty 100 -intron_penalty 130” as suggested by (Zhang and Gish 2006). Default parameters and “A=4” were used for Sim4 to output alignments. PAIRAGON and QPAIRAGON parameters were estimated from the EST_GENOME alignments on Set B first. In order to avoid training and test on the same data, a 4-fold cross-validation was performed for both PAIRAGON and QPAIRAGON. All EST alignment clusters in Set B were divided into four random groups. EST_GENOME alignments of three of the four groups were picked as the training set. The estimated parameters were used to run on the remaining group of Set B. This procedure was repeated four times so that each group was in the test set once.

5.10.1 Result Comparison on Explained Mismatches

The new accuracy measurement described in Section 5.8 was assessed for each program. All programs were evaluated on the alignments for human chromosome 20, 21 and 22 of Set B. Table 5-3 shows the alignment result comparison based on the new accuracy measurement. The SNP database was downloaded from UCSC on July 15, 2005. There are 38,509 SNPs in coding regions or UTRs of human chromosome 20, 21 and 22. QPAIRAGON performs as well in both sensitivity and specificity as EST_GENOME. Sim4 did not fit for this data set. A lot of the times, it did not produce an alignment result.

Table 5-3. Alignment result comparison by explained number of SNPs.

Methods	Aligned Mis-matches	SNPs	Explained SNPs	Sensitivity	Specificity
EST_GENOME	37,360	38,509	1,569	0.041	0.042
Sim4	25,828	38,509	1,109	0.029	0.043
PAIRAGON	28,647	38,509	1,471	0.038	0.051
QPAIRAGON	37,438	38,509	1,573	0.041	0.042

5.10.2 Accuracy on Different Quality Values

Similar to mismatches explained by SNPs, a mismatch may be called “explained by a low quality value” if the quality value of the EST base is lower than a threshold. Table 5-4 shows the numbers of explained mismatches with quality value lower than threshold 5, 10, 15, and 20 for EST_GENOME, Sim4, PAIRAGON and QPAIRAGON on chromosome 20, 21 and 22 of Set B. Compared to other methods, significantly more may be caused by EST bases with low quality values. Therefore, a mismatch aligned by QPAIRAGON is more reliable than a mismatch by EST_GENOME.

Table 5-4. Mismatch numbers versus quality values.

From left to right, the second column is for the absolute number of mismatches for each method. The third column is for mismatches that can be explained by SNPs. The rest columns are for mismatches that can be explained by quality values lower than 5, 10, 15 and 20. From the third column, there are two rows for each method. The upper row is for the absolute numbers and the lower row is for the portion of total mismatches.

	Total	SNPs	5	10	15	20
EST_GENOME	37,360	1,569	6,737	14,322	16,653	17,861
		0.042	0.180	0.383	0.446	0.478
Sim4	25,828	1,109	3,595	9,114	10,945	12,025
		0.043	0.139	0.353	0.424	0.466
PAIRAGON	28,647	1471	4976	10789	12683	13717
		0.051	0.161	0.377	0.443	0.479
QPAIRAGON	37,464	1,574	7,198	16,737	19,196	20,331
		0.042	0.192	0.447	0.512	0.543

The expected number of mismatches for each method was counted by weighting each mismatch with the probability associated with the sequencing quality value. Table 5-5 shows that the expected numbers of mismatches explained by SNPs for QPAIRAGON and EST_GENOME are similar, but QPAIRAGON has lower expected number of mismatches than EST_GENOME. PAIRAGON has much lower expected number of mismatches than QPAIRAGON and EST_GENOME, but its number of explained mismatches is lower too.

Table 5-5. Expected number of mismatches.

From left to right, the first column is for the total number of aligned mismatches; the second column is for the expected total number of aligned mismatches; the third column is for the expected number of mismatches explained by SNPs. The expected sensitivity was computed by comparing the expected number of explained mismatches to the total number of SNPs, which is 38,509. The expected specificity was computed by comparing the expected number of explained mismatches to the expected number of mismatches in the first column.

Methods	Aligned Mis-matches	Expected Aligned Mismatches	Expected Explained SNPs	Expected Sensitivity	Expected Specificity
EST_GENOME	37,360	29,665	1,511	0.0392	0.0509
Sim4	25,828	21,445	1,067	0.0277	0.0498
PAIRAGON	28,647	23,077	1,421	0.0369	0.0616
QPAIRAGON	37,438	29,030	1,513	0.0393	0.0521

5.10.3 Result Comparison on Exon-intron Level

By using the human RefSeq annotations as the standard, accuracy of finding the exact exon-intron boundaries was compared for EST_GENOME, PAIRAGON and QPAIRAGON. The results are shown in Table 5-6. RefSeq annotation on human chromosome 20, 21 and 22 were used as the annotation. All the aligned regions of EST alignments were treated as coding regions. Since one EST only covers part of a gene and many EST reads aligned to UTRs, this may explains the low sensitivity and specificity compared to those alignment results on high quality cDNA to genome alignments. Many EST contain polyA regions at their ends though they may from the universal 3’ primers for reverse-transcription. These polyA regions can be aligned to many positions with multiple continuous “A”s in a genome sequences and produce extra false introns. For all three methods, an exon was excluded if it is either the first or the last exon and at least 60% of bases were all “A”s or all “T” alone.

Table 5-6. Exon-Intron boundary finding accuracy comparison.
The sensitivity and specificity are based on RefSeq coding region annotation on chromosome 20 21 and 22. The real accuracy may be higher if UTR exons and introns are considered.

	Exon_Sn	Exon_Sp	Intron_Sn	Intron_Sp
EST_GENOME	21.3	27.8	31.3	68.6
PAIRGON	20.7	27.0	30.8	66.9
QPAIRAGON	21.6	27.9	31.9	68.2

5.11 Gene Prediction Pipeline with the New Alignment Program

One of the important goals of developing a new spliced alignment program in this thesis is to improve gene prediction accuracy. BLAT has been the default EST-to-genome alignment program in the gene prediction pipeline we developed to use ESTs. Different EST-to-genome alignment programs, such as EST_GENOME and QPAIRAGON, were plugged into TWINSCAN_EST prediction pipeline. All ESTs from Set A were used to create ESTseqs for human genome chromosome 20, 21 and 22. Default parameters were used to create the BLAT alignment and parameters for EST_GENOME were the same as in the previous section. In order to avoid training and testing on the same data set, a 4-fold cross-validation was used to create QPAIRAGON alignments. The parameters estimated from three groups were tested on the remaining one group, except the test group contained all the ESTs instead of just the non-overlapping ESTs as in the previous section. After the EST-to-genome alignments were created by each spliced-alignment program, ESTseqs for human chromosome 20, 21 and 22 (version NCBI35) were created. For each alignment method, an 8-fold cross validation procedure was performed for TWINSCAN_EST. With RefSeq annotation on chromosome 20, 21 and 22 as standard, the ESTseq parameters of TWINSCAN_EST were estimated from seven groups and tested on the remaining group. In this way, the training and testing on the same data problem was avoided for all alignment programs. Finally, the overall gene prediction results were evaluated on all 8 folds together. The results shown in Table 5-7 demonstrate that QPAIRAGON is the most effective one to use the EST information. It performs better throughout all measurements, especially in gene level. It obtained 1% more accurate in gene sensitivity than EST_GENOME and about 2% more accurate in both gene sensitivity and specificity than BLAT did. These improvements indicate that alignment quality of QPAIRAGON is higher than EST_GENOME and BLAT.

Table 5-7. Effect of Underlying Alignment Programs on TWINSCAN_EST Gene Prediction. ESTseqs were generated by using BLAT, EST_GENOME and QPAIRAGON as the underlying alignment program in TWINSCAN_EST system. Then TWINSCAN_EST was run on human chromosome 20, 21 and 22 with ESTseqs generated by these alignment programs. The sensitivity and specificity is based on the 897 RefSeq genes in chromosome 20, 21 and 22.

	Gene_Sn	Gene_Sp	Exon_Sn	Exon_Sp
BLAT	33.6	19.4	75.0	62.5
EST_GENOME	34.9	20.5	76.5	62.5
QPAIRAGON	35.8	21.1	76.6	62.8

5.12 QPAIRAGON on Cross-Species Alignment

Experiments similar to those discussed in Section 5.11 were conducted for cross-species alignment. BLAT, EST_GENOME and QPAIRAGON were used to align mouse genome (version mm6) with all the human EST reads described in Section 5.9. By using BLAT as the alignment tool, there are 273, 262, 158, 125 and 134 human EST clusters aligned to mouse chromosome 1, 5, 17, 18 and 19 respectively. The total numbers of aligned ESTs with at least 95% matches on an aligned region are 679, 648, 626, 323 and 316 for mouse chromosome 1, 5, 17 18 and 19 respectively. These numbers are smaller than those generated from native EST-to-genome alignments. TWINSCAN_EST was run on these five mouse chromosomes using BLAT, EST_GENOME and QPAIRAGON as the underlying spliced alignment program to create ESTseqs. Human genome NCBI35 was the informant database. Default parameters were used for EST_GENOME alignments. The QPAIRAGON parameters were estimated from EST_GENOME alignments of ESTs to five human chromosomes. For each alignment method, only those alignments with matches for at least 95% of an aligned region were kept for ESTseq generating. Result in Table 5-8 show that the three spliced alignment programs have very similar performance though QPAIRAGON performs slightly better throughout all evaluation categories.

Table 5-8. The effect of spliced alignment programs on gene prediction with ESTseqs created by cross-species EST to genome alignment.

BLAT, EST_GENOME and QPAIRAGON were used to generate ESTseqs for 5 mouse chromosomes (mm6). TWINSCAN_EST was run with human genome NCBI35 as informant database. Accuracy evaluation is based on 2577 genes and 25177 exons annotated for 5 mouse chromosomes.

	Gene_Sn	Gene_Sp	Exon_Sn	Exon_Sp
BLAT	22.8	10.3	66.0	45.2
EST_GENOME	22.9	10.3	65.9	45.2
QPAIRAGON	23.0	10.4	66.0	45.3

5.13 A QPAIRAGON Alignment Example

Figure 5-9 shows a QPAIRAGON alignment example. An EST with id 154078196 was aligned to human chromosome 21. The sequencing quality values of the EST were shown along the optimal alignment. EST_GENOME alignment of this EST missed the whole intron and the right region flanking this intron since GC/AG intron is not preferred by EST_GENOME. Interestingly, PAIRAGON also predicted an intron but it was a GT/AG intron (See Figure 5-10). Since all bases around EST position 527 are in high quality, insertion or deletion has much higher penalty than substitutions for QPAIRAGON. Although the GC/AG intron is much less likely, it is preferred in this case. One question is which alignment should be trusted? In order to answer this question precisely, Section 5.14 will introduce the posterior probability for all QPAIRAGON alignment.

```
   chr21          10494 GTATGAAAGATCTAATTTCTCTACGGCctcac......actgcACTCTAG 18940
                        |||||||||||||||||||||||| ||<<<<< 8413 <<<<<|||| ||
Ti|154078196        501 GTATGAAAGATCTAATTTCTCTACTGC................ACTCCAG   534
                        333343432222333444422243334                4455544
                        655503873333259562299923570                8266622

   chr21          18941 CCTGGGTGACAGAGTGAGACTC--TCAAAAAAAACAAAAACAAAAAAACA 18988
                        ||| ||  |||||| |||||||  || ||||||   |||| ||||||| |
Ti|154078196        535 CCTAGGCAACAGAGCGAGACTCCGTCGAAAAAAGGGAAAA-AAAAAAA-A   582
                        4445433333444443323333444433342222122444 4444455 5
                        7726077775227224487044000077685995905088 8888866 6
```

Figure 5-9. A QPAIRAGON alignment example.

An EST with id 154078196 was aligned to human chromosome 21. The chromosome coordinates are relative coordinates on a 20,432 base region [13,351,011, 13,371,528] of chromosome 21. The sequencing quality values of the EST were shown along the optimal alignment. The alignment is divided into 50 base width fragments. Within each fragment, the first line is a genomic sequence from human genome chromosome 21; the third line is the EST sequence and the second line is the alignment sequence between the genomic sequence and EST sequence. The next two lines are for the sequencing quality values. For example, the first numbers from the fourth and fifth line of the first fragment means that the quality value at position 451 is 37.

```
    chr21    10494 GTATGAAAGATCTAATTTCTCTACGGCctcac......gccacTGCACTC  18937
                   ||||||||||||||||||||||||   <<<<< 8410 <<<<<|||||||
ti|154078196   501 GTATGAAAGATCTAATTTCTCTAC---................TGCACTC    531

    chr21    18938 TAGCCTGGGTGACAGAGTGAGACTC--TCAAAAAAA---ACAAAAACAAA  18982
                    ||||| ||  |||||| |||||||  || ||||||   | ||||| |||
ti|154078196   532 CAGCCTAGGCAACAGAGCGAGACTCCGTCGAAAAAAGGGAAAAAAAAAAA    581
```

Figure 5-10. A PAIRAGON alignment example. Both genomic sequence and EST sequence are the same as in the QPAIRAGON example.

5.14 Alignment Quality of QPAIRAGON

A question naturally arises when an alignment is created. The question is: How reliable is the alignment? Are there any similar alignments other than this one? Posterior probability may be used to quantify the degree of uncertainty on each aligned base pair upon all alignment paths. QPAIRAGON can be instructed to generate posterior probability in two different ways:

1) Posterior probability of a state, "Match/Mismatch" for example, can be calculated for each alignment pair (x_i, y_j), where x_i is a base from the genomic sequence and y_j is a base from the EST sequence.

2) The posterior probability of a state, "Match/Mismatch" for example, can be output along the optimal alignment.

The output of the first method may be very large, so it can be used for each individual alignment. The second method can give you the reliability of the optimal alignment on each EST base. When posterior probabilities of the "Match/Mismatch" state are output along the optimal path, each aligned base pair in an exon part of the optimal alignment can be evaluated by the posterior probability of that base pair as a match or mismatch. Figure 5-11 shows an example of using posterior probability of the "Match/Mismatch" state as the measurement of uncertainty. Both genomic sequence and the EST sequence are the same as those used in the example shown in the previous section. The "GC" at genomic position [18333, 18332] is clearly not in a match state according to the posterior probability, while the "AG" at genomic position [10520, 10519] may be only an alternative acceptor site because there are more "AG"s at position [10515, 10514] and [10513, 10512]. Based on the posterior probability, we can say QPAIRAGON alignment is more reliable than PAIRAGON alignment in this case.

The posterior probability in genomic position 10514 is lower than its neighbors'. This may be caused by the two-base-long states next to the intron state to model the splice donor and acceptor sites in the genomic sequence. Because the scores are only stored in the one base of the two bases. This can be fixed by splitting the two-base-long states into two one-base states or switching the HMM to a GHMM though posterior probability in positions other than potential splice sites is not likely to be affected by this.

```
                   99999999999999999999891100000000......000009999999
                   99999999999999999999271011000000......000009999999
   chr21     10494 GTATGAAAGATCTAATTTCTCTACGGCctcac......actgcACTCTAG 18940
                   |||||||||||||||||||||||| ||<<<<< 8413 <<<<<|||| ||
ti|154078196   501 GTATGAAAGATCTAATTTCTCTACTGC................ACTCCAG   534
                   333343432222333444422243334                4455544
                   655503873333259562299923570                8266622

                   99999999999999999999996099988877775333331111111110
                   99995766999999999999999509962118531781111598999919
   chr21     18941 CCTGGGTGACAGAGTGAGACTC--TCAAAAAAAACAAAAACAAAAAAACA 18988
                   ||| ||  |||||| |||||||  || ||||||   |||| ||||||| |
ti|154078196   535 CCTAGGCAACAGAGCGAGACTCCGTCGAAAAAAGGGAAAA-AAAAAAA-A   582
                   4445433333444443323333444433342222122444 4444455 5
                   7726077775227224487044000077685995905088 8888866 6
```

Figure 5-11. A QPAIRAGON alignment example with posterior probability of match/mismatch displayed along the optimal path.

The number of match/mismatches with a certain posterior probability was counted for each level of posterior probability on alignment results of Set B on human chromosome 20, 21 and 22. The posterior probability here is the posterior probability with constraints used by stepping stone algorithm. Figure 5-12 shows that the optimal path can be either a dominating path with posterior probability larger than 0.95 or one of many possible paths with posterior probabilities close to zero. For QPAIRAGON, the posterior probability depends more on the sequences than the sequencing quality values. A low posterior probability on match state often implies another possible splicing site nearby.

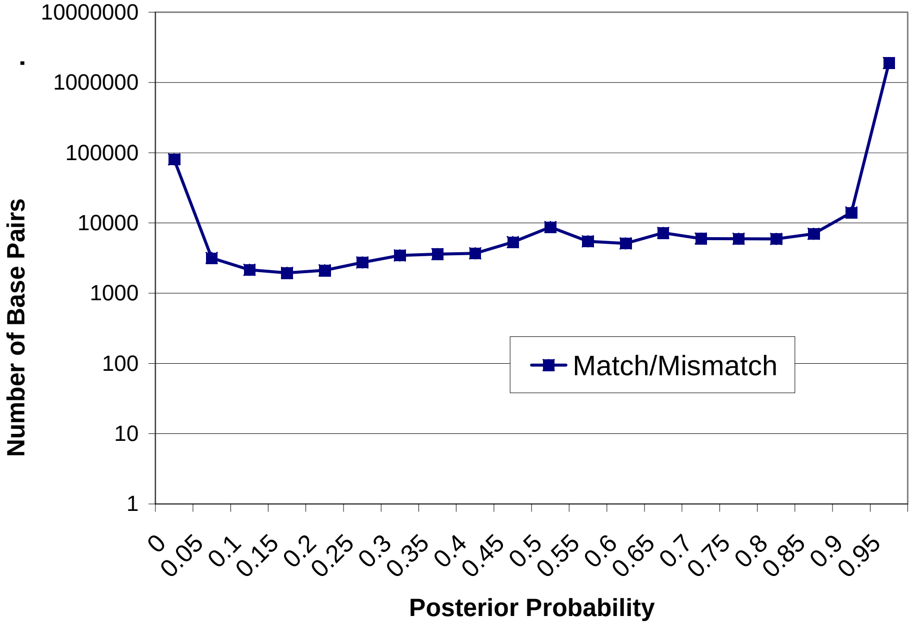


Figure 5-12. Posterior probability distribution for Match/Mismatch state.
The x-axis is for posterior probability and y-axis is for the number of base pairs in Match/Mismatch state for each level of posterior probability.

5.15 Conclusions and Discussions

This Chapter introduced QPAIRAGON, a new EST-to-genome sequence alignment system. By using the EST sequencing quality values, it can align EST to genome more accurately. Compared to EST_GENOME, the existing most accurate spliced alignment system in most situations, QPAIRAGON can align intron-exon boundaries more accurately and more mismatches can be expected to be explained by known

SNPs. The accuracy of gene prediction on human genome can be improved about one to two percents when it is used for comparative-genomics-based novel gene prediction systems, such as TWINSCAN_EST.

The gene prediction accuracy improvement by using ESTs from different organism is marginal for all the spliced alignment programs tested. One of the reasons is that the high identity percentage level in pre-processing procedure filtered out many ESTs when they are aligned to a different genome. However, the gene prediction accuracy with QPAIRAGON as the underlying alignment program is at least as good as other alignment programs, such as BLAT and EST_GENOME. Therefore, it's safe to say that QPAIRAGON is as effective as other spliced alignment systems when aligning ESTs to a genome of different organism.

QPAIRAGON requires EST sequencing quality values in addition to the EST and genomic sequences. Sometimes, the quality values were not submitted to databases, especially for those ESTs generated a certain years before. Another disadvantage of the system is that the accuracy improvement comes with a trade off in speed. However, with the two-stage strategy, we demonstrated that it can be used to deal with thousands of ESTs in a reasonable time. It can be useful for applications that require high accuracy alignments, such as manual curation of gene structure and SNP finding.

Chapter 6

Conclusions

The previous Chapters showed the probability models of using ESTs to improve de novo gene predictions, using sequencing quality value to improve the EST-to-genome alignment, and using consistent gene models of an EST alignment to guide full ORF-containing-cDNA finding.

One of the most important progresses in gene prediction in the past years is the success of comparative-genomics-based gene prediction systems. The main accomplishment of this study was the development of new gene prediction programs that can integrate information from EST alignments with comparative-genomics-based gene predictors. The new gene prediction systems are able to improve the accuracy (both sensitivity and specificity) of gene prediction on genes that have aligned ESTs. Their accuracy in predicting complete gene structures is as good as that of the original, non-EST-aware programs, even when no ESTs aligned to the target genome. When genes with no aligned ESTs are interspersed with genes that have aligned ESTs, the accuracy on the genes without ESTs is higher than that of the original programs. Another notable feature of this approach to use EST alignments is that every aligned EST bases contribute to gene prediction as long as the overall EST alignment meets a certain quality criterion. ESTs regions aligned to UTR region can improve the UTR region prediction as well as the coding region prediction.

Therefore, the use of EST information is very effective and comes at no cost. TWINSCAN_EST and N-SCAN_EST have the essential benefit of a de novo gene finder, the ability to find completely novel genes without sequence similarity to known genes. And they are more accurate on genes for which ESTs are available. In particular, TWINSCAN_EST can predict the 75% of genes exactly for *C. elegans* genome and 72% of the predicted genes are exactly right. N-SCAN_EST can predict 44% of genes exactly and 88% of exons in human genome. Compared to other de novo gene finders, TWINSCAN_EST is the most accurate program available for nematodes and N-SCAN_EST is the most accurate program available for mammals. This method to use ESTs can be applied to gene prediction for any organism. It is especially useful for newly sequenced genomes for which some ESTs are available but few full-length cDNAs have been generated. The extension of this method to a new genome is pretty straightforward and re-training the parameters from native EST alignments can improve the accuracy slightly further. Since the use of EST information comes at no cost, it is recommended to use the EST versions of these programs on any genomes, with or without ESTs.

Another advantage of the novel gene prediction programs using EST information is their ability to predict UTR regions more precisely since an EST contains coding regions or UTR regions or both of them. The improvement in predicting transcription boundary might be helpful to predict signals near them, such as transcription promoters.

How far can this method of using EST alignments for gene prediction go? In other words, what is the best performance can this method produce? For example, what will happen if the error prone ESTs are replaced with high quality full-length cDNAs? Experiments show that the performance of gene prediction systems with EST alignments could not predict the genes as accurately as those methods purely based on alignments. Therefore, it is better to use alignment-based method for those high quality transcript product resources like proteins or full-length cDNAs and use this approach to deal with ESTs.

CDNA genomic sequence alignment is critical to many computational biology applications. A new cDNA-to-genome alignment program QPAIRAGON is introduced to use sequencing quality values to improve the alignment accuracy. A graphical model is also created to represent the error patterns arising from sequencing error and polymorphism (or evolutionary divergence if the ESTs and genomic sequences are from different organisms). By using an alignment scoring system that takes quality values into account, the new alignment program can produce more accurate EST-to-genome alignments than the best existing cDNA-to-genome alignment programs. It makes the EST resources more valuable for computational biology applications like single nucleotide polymorphisms (SNPs) and alternative splicing detection, the success of which is heavily depended on the alignment quality. The posterior probability function of the QPAIRAGON can be very useful to determine some ambiguous alignments. It can be helpful to resolve some difficult cases when ESTs are used in manual gene structure annotation.

The graphical model created in Chapter 5 built up a framework to include the sequencing error and polymorphism as well as evolutionary divergence if the ESTs and genomic sequences are from different organism. Though the accuracy of cross-species spliced alignment by QPAIRAGON is not significantly better than EST_GENOME, this remains a promising direction for future work.

Due to the speed trade off, we are not recommending QPAIRAGON to replace those fast heuristic based spliced alignment programs now. QPAIRAGON can be speed up by applying more stringent boundary constraints the stepping stone algorithm currently used.

Appendix A

Biology Background

Appendix A provides the necessary biology background to understand the motivation and methods of this research.

Genome and gene

All the genetic information is encoded in deoxyribonucleic acid (DNA). A genome is all DNAs of an organism while a gene is a relatively short piece of DNA that represents a fundamental physical and functional unit of heredity. A DNA sequence is consisted of four bases: the purines adenine (A) and guanosine (G) and the pyrimidines cytosine (C) and thymine (T). The A pairs to T and C pairs to G. The size of a genome can be represented by the number of base pairs it contains. For example, the human genome has about 3 billion base pairs (bps), and about 25,000 protein coding genes. In this dissertation, since we are only interested in protein coding genes, a “gene” means a protein coding gene.

Figure 1 shows a diagram of the structure of a gene. The three important features in a gene structure are untranslated regions (UTRs), coding regions (exons) and introns (regions between two adjacent coding regions). The usage of the term “Exon” varies in different literatures. In this dissertation it stands for the coding region exon. It can also be used to include UTRs. When it is used for UTRs, it is called “UTR exons” in this dissertation. The first two bases (5’ end) of an intron can be used to represent the pattern of the splice donor site, which is the region around the 5’ end of the intron. Similarly, the last two bases (3’ end) of an intron can be used to represent the pattern of the splice acceptor site, which is the region around the 3’end of an intron. Most of the times, a splice donor site pattern is GT and a splice acceptor site pattern is AG. An intron is called a canonical intron if its 5’ end is GT or GC and its 3’ end is AG. A real gene structure can be much more complicated than the one shown in this figure. Based on N-SCAN_EST’s prediction on the human genome, the average length of a human gene is about 46,000 bases, and there are about 8 exons per gene on average. The human gene intron lengths (4,500 bases on average) are much longer than exon lengths (145 bases on average for internal exons) and the average total length of coding region for a human gene is about 1,300 bases. Therefore the total protein coding region is only a tiny part of the 3 billion bases in the whole human genome (about 1%). These numbers are different for different organisms. For example, intron length of the *C. elegans* (a worm) gene is much shorter, about 280 bases on average, while the average coding region length is about 1,160 bases, which is relatively close to human coding region length.

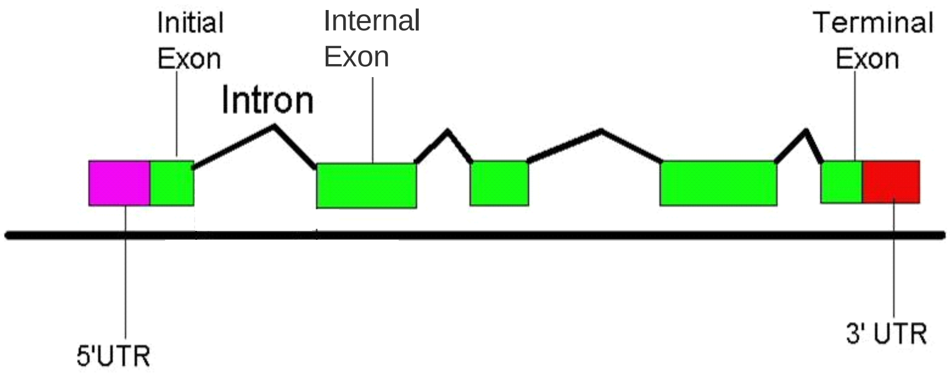


Figure 1. A simple diagram of a protein coding gene structure.

The left side purple box is 5’ UTR, green boxes are coding regions (exons), the right side red box is 3’UTR and a region between two adjacent green boxes is an intron region. The number of introns in a gene can be zero, which means there is only one exon, called a single exon and the gene is called a single-exon gene. The first codon (3 nucleotides) of an initial exon (or a single exon), where the translation starts, is ATG; and the last codon of a terminal exon (or a single exon), where the translation stops, is one of the three codons TAA, TAG and TGA. At each end of an intron region is a splice site: 5’ end is a splice donor site and 3’ end is a splice acceptor site.

DNA, mRNA, Transcription, Translation and Alternative Splicing

The central dogma of molecular biology describes the flow of the information in biology from DNA to mRNA and finally to protein. Information stored in DNAs is transferred to mRNAs by using DNAs as the templates for mRNA polymer synthesis. Proteins are made by using mRNAs as the templates. In the procedure from a DNA to an mRNA, the DNA sequence is transcribed into a pre-mature mRNA, and at the same time a procedure called splicing removes the introns and concatenates the remaining contiguous regions together as an mRNA. Therefore, an mRNA contains no intron regions, which means only UTRs and coding regions of a gene are included in an mRNA polymer. During the splicing procedure, multiple alternative splice donor sites can be joined to a common splice acceptor site and multiple splice acceptor sites can be joined to a common splice donor site. In this way, different splice donor and/or acceptor sites can be used to produce multiple mRNAs with partially overlapping genetic information. This is called alternative splicing. Each of these mRNAs is called a transcript of the gene and each transcript mat be translated into a distinct protein. Alternative splicing is very common in mammalian genomes, and about 30% to 60 % of human genes are estimated to have distinct alternative splicing (Modrek, Resch et al. 2001; Modrek and Lee 2002; Lee, Atanelov et al. 2003; Modrek and Lee 2003).

Exon, Frame and Translation

The usage of the term “Exon” varies in different literatures. In this dissertation it stands for the coding region. It can also be used to include UTRs. When it is used for UTRs, it is called “UTR exons” in this dissertation. The nucleotide sequence of an mRNA is translated to a sequence of amino acids. Each amino acid is determined by three continuous nucleotides called a codon. Thus the coding region length of a gene is always a multiple of 3. In animals, translation always starts from a codon “ATG”, which is called the start codon. The boundary of an internal coding exon can start in the middle of a codon. In another words, starting from the first exon, if an exon has a length other than a multiple of 3, then the following exon will start in a different position of a codon. The codon position in which an exon starts with is called the frame or phase of the exon. The codon usage of exons with different frames has been used to improve gene structure prediction.

CDNA, cDNA Clone, and cDNA library

A complementary DNA (cDNA) is a form of DNA prepared in a laboratory with an mRNA as the template by a procedure called reverse transcription (from an mRNA to a DNA). Therefore, as an mRNA, a cDNA contains only UTRs and/or coding regions of its gene. An Open Reading Frame (ORF) is a portion of a gene's sequence that is uninterrupted by a stop codon so that it encode a peptide or protein. A full-ORF is an ORF containing all coding region of a gene. A full-ORF cDNA is a cDNA that contains all the coding region of a gene, from the start of translation to the end of translation. A full-length cDNA is a cDNA that goes from the end of transcription to the start of transcription. Full-length cDNAs are extremely useful for determining the gene structures of a genome. The most direct experimental evidence of a gene structure is from sequencing a full-length cDNA and aligning its sequence back to the genomic sequence.

A clone is collection of genetically identical cells which are generated from the same cell. A cloning vector is a bacterial virus or plasmid which can contain one or more segments of foreign DNAs and without loss of the capability of self-replication in a host bacterium. Many genetically identical cloning vectors can be generated for each cloning vectors in its host bacterium, which become a clone. Full-ORF cDNA clones are clones containing full-ORF cDNA segments. They are very important resources for functional genomics researches since they contain the whole gene structures and can be stored and used very conveniently.

A cDNA library is a set of all or nearly all the mRNAs contained within a cell or organism. Since mRNA is not stable, cDNAs of the mRNAs were produced by reverse transcription. The set of cDNAs, thus the mRNAs, is collectively called a cDNA library.

Random-cDNA-Clone Selection

In biology experiments, many clones are grown in each plate. It is not known which clone contains a full-length cDNA. Different strategies have been developed for full ORF-containing-cDNA-clone selection. One of them is randomly pick a clone and sequence the cDNA sequence to high quality, which is an expensive and time consuming procedure. Improvement of this random selection includes the use of ESTs, which are from either or both end of cDNAs. A decision is made based on the result of ESTs. If the ESTs indicate that this cDNA is highly possible to contain a full-ORF, the cDNA is sequenced deeply to get high quality cDNA sequences with a high priority.

EST

Expressed Sequence Tags (ESTs) are generated by single sequencing reads from either one or both ends of cDNA clones (Adams, Kelley et al. 1991; Boguski, Lowe et al. 1993; Schultz, Doerks et al. 2000). A 5' EST is sequenced from 5' end of a cDNA clone, and a 3' EST is sequenced from 3' end of a cDNA clone. Since they are single reads, they are short, usually with a length of 200 to 500 reliable bases, which means a single EST can only cover a small portion (coding regions or/and UTRs) of a gene; and their sequencing error rates are very high, about 3% compared to the error rate of 0.01% for genomic sequences produced by most large scale genome sequencing projects. Nevertheless, the advantage of ESTs is that they can be generated fast and inexpensively (Adams, Kelley et al. 1991; Boguski, Lowe et al. 1993; Schultz, Doerks et al. 2000; Kan, Rouchka et al. 2001). dbEST (Boguski, Lowe et al. 1993; Boguski 1995) is the EST division of GenBank. As of January 20, 2006, dbEST had about 33 million ESTs for 1021 organisms, of which about 7.6 million are human ESTs and about 4.7 million are mouse ESTs (http://www.ncbi.nlm.nih.gov/dbEST). They provide tremendous resources for genomic sequence analysis.

Appendix B

dbEST Accession Numbers for Experiment Sequences

The Trace Archive and dbEST Accession numbers for experiment sequences to validate N-SCAN_EST partial novel predictions. An experiment sequence is included here if its best alignment met the following criteria:

1. It contained at least 50 consecutive matches,
2. At least 95% of the entire sequence were matches,
3. It contained at least one intron.

DR731296 DR731297 DR731298 DR731300 DR731302
DR731303 DR731306 DR731307 DR731309 DR731311
DR731313 DR731316 DR731317 DR731318 DR731321
DR731322 DR731323 DR731324 DR731325 DR731326
DR731327 DR731328 DR731333 DR731334 DR731335
DR731336 DR731337 DR731338 DR731339 DR731340
DR731341 DR731342 DR731343 DR731344 DR731345
DR731346 DR731347 DR731348 DR731351 DR731352
DR731353 DR731354 DR731355 DR731356 DR731361
DR731362 DR731367 DR731368 DR731375 DR731376
DR731380 DR731383 DR731384 DR731387 DR731388
DR731391 DR731392 DR731393 DR731394 DR731395
DR731396 DR731397 DR731398 DR731399 DR731400
DR731403 DR731404 DR731405 DR731406 DT932517

DT932519 DT932520 DT932523 DT932525 DT932526
DT932527 DT932528 DT932529 DT932530 DT932531
DT932532 DT932533 DT932534 DT932535 DT932536
DT932537 DT932539 DT932540 DT932541 DT932542
DT932543 DT932550 DT932551 DT932552 DT932553
DT932555 DT932559 DT932560 DT932566 DT932567
DT932568 DT932571 DT932572 DT932573 DT932574
DT932577 DT932578 DT932587 DT932588 DT932589
DT932590 DT932591 DT932592 DT932593 DT932594
DT932603 DT932604 DT932607 DT932608 DT932609
DT932610 DT932613 DT932614 DT932615 DT932616
DT932617 DT932618 DT932619 DT932620 DT932623
DT932624 DT932625 DT932626 DT932629 DT932630
DT932631 DT932632 DT932633 DT932634 DT932635
DT932636 DT932639 DT932640 DT932641 DT932642
DT932643 DT932644 DT932645 DT932646 DT932649
DT932650 DT932653 DT932654 DT932659 DT932660
DT932661 DT932662 DT932663 DT932664 DT932665
DT932666 DT932667 DT932668 DT932669 DT932670
DT932673 DT932674 DT932675 DT932676 DT932677
DT932678 DT932683 DT932684 DT932693 DT932694
DT932695 DT932696 DT932701 DT932702 DT932709
DT932710 DT932715 DT932716 DT932719 DT932720
DT932721 DT932722 DT932723 DT932724 DT932725
DT932726 DT932727 DT932728 DT932731 DT932732
DT932735 DT932736 DT932739 DT932740 DT932743
DT932744 DT932745 DT932746 DT932749 DT932750
DT932751 DT932752 DT932753 DT932754 DT932761
DT932762

References

Adams, M. D., J. M. Kelley, et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science **252**(5013): 1651-6.

Adams, M. D., Kelley, J.M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R.F, et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science **252**(5013): 1651-6.

Allen, J. E., M. Pertea, et al. (2004). "Computational gene prediction using multiple sources of evidence." Genome Res **14**(1): 142-8.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Ashburner, M. (2000). "A biologist's view of the Drosophila genome annotation assessment project." Genome Res **10**(4): 391-3.

Birney, E., M. Clamp, et al. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-95.

Boguski, M. S. (1995). "The turning point in genome research." Trends Biochem Sci **20**(8): 295-6.

Boguski, M. S., T. M. Lowe, et al. (1993). "dbEST--database for "expressed sequence tags"." Nat Genet **4**(4): 332-3.

Brendel, V., L. Xing, et al. (2004). "Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus." Bioinformatics **20**(7): 1157-69.

Brent, M. R. (2005). "Genome annotation past, present, and future: how to define an ORF at each locus." Genome Res **15**(12): 1777-86.

Brent, M. R. and R. Guigo (2004). "Recent advances in gene structure prediction." Curr Opin Struct Biol **14**(3): 264-72.

Brown, R. H., S. S. Gross, et al. (2005). "Begin at the beginning: predicting genes with 5' UTRs." Genome Res **15**(5): 742-7.

Burge, C. (1997). Identification of genes in human genomic DNA, Stanford University.

Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Burset, M. and R. Guigo (1996). "Evaluation of gene structure prediction programs." Genomics **34**(3): 353-67.

Churchill, G. A. (1989). "Stochastic models for heterogeneous DNA sequences." Bull Math Biol **51**(1): 79-94.

Durbin, R., Eddy,S., Krogh, A., Mitchison, G. (1998). Biological sequence analysis, Cambridge University.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-94.

Ewing, B., L. Hillier, et al. (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-85.

Fickett, J. W. (1982). "Recognition of protein coding regions in DNA sequences." Nucleic Acids Res **10**(17): 5303-18.

Fields, C. A. and C. A. Soderlund (1990). "gm: a practical tool for automating DNA sequence analysis." Comput Appl Biosci **6**(3): 263-70.

Flicek, P., E. Keibler, et al. (2003). "Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map." Genome Res **13**(1): 46-54.

Flicek, P., E. Keibler, et al. (2003). "Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map." Genome Res **13**: 46-54.

Florea, L., G. Hartzell, et al. (1998). "A computer program for aligning a cDNA sequence with a genomic DNA sequence." Genome Res **8**(9): 967-74.

Gelfand, M. S. (1990). "Computer prediction of the exon-intron structure of mammalian pre-mRNAs." Nucleic Acids Res **18**(19): 5865-9.

Gibbs, R. A., G. M. Weinstock, et al. (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." Nature **428**(6982): 493-521.

Gross, S. S. and M. R. Brent (2005). Using Multiple Alignments To Improve Gene

Prediction. RECOMB, Boston.

Guigó, R., P. Agarwal, et al. (2000). "An assessment of gene prediction accuracy in large DNA sequences." Genome Res **10**(10): 1631-42.

Guigó, R., E. T. Dermitzakis, et al. (2003). "Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes." Proc Natl Acad Sci U S A **100**: 1140-1145.

Guigo, R., E. T. Dermitzakis, et al. (2003). "Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes." Proc Natl Acad Sci U S A **100**(3): 1140-5.

Guigó, R., S. Knudsen, et al. (1992). "Prediction of gene structure." J Mol Biol **226**(1): 141-57.

Guigo, R. and M. G. Reese (2005). "EGASP: collaboration through competition to find human genes." Nat Methods **2**(8): 575-7.

Harris, T. W., N. Chen, et al. (2004). "WormBase: a multi-species resource for nematode biology and genomics." Nucleic Acids Res **32 Database issue**: D411-7.

Harris, T. W., R. Lee, et al. (2003). "WormBase: a cross-species database for comparative genomics." Nucleic Acids Res **31**(1): 133-7.

Hillier, L. W., W. Miller, et al. (2004). "Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution." Nature **432**(7018): 695-716.

Howe, K. L., T. Chothia, et al. (2002). "GAZE: a generic framework for the integration of gene-prediction data by dynamic programming." Genome Res **12**(9): 1418-27.

Huang, X., M. D. Adams, et al. (1997). "A tool for analyzing and annotating genomic sequences." Genomics **46**(1): 37-45.

Kan, Z., E. C. Rouchka, et al. (2001). "Gene structure prediction and alternative splicing analysis using genomically aligned ESTs." Genome Res **11**(5): 889-900.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-64.

Korf, I., P. Flicek, et al. (2001). "Integrating genomic homology into gene structure prediction." Bioinformatics **17 Suppl 1**: S140-8.

Krogh, A. (2000). "Using database matches with for HMMGene for automated gene detection in Drosophila." Genome Res **10**(4): 523-8.

Krogh, A., M. Brown, et al. (1994). "Hidden Markov models in computational biology. Applications to protein modeling." J Mol Biol **235**(5): 1501-31.

Krogh, A., I. S. Mian, et al. (1994). "A hidden Markov model that finds genes in E. coli DNA." Nucleic Acids Res **22**(22): 4768-78.

Kulp, D., D. Haussler, et al. (1996). "A generalized hidden Markov model for the recognition of human genes in DNA." Proc Int Conf Intell Syst Mol Biol **4**: 134-42.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lauritzen, S. L. (1996). Graphical models. Oxford, Clarendon Press.

Lee, C., L. Atanelov, et al. (2003). "ASAP: the Alternative Splicing Annotation Project." Nucleic Acids Res **31**(1): 101-5.

Maglott, D. R., K. S. Katz, et al. (2000). "NCBI's LocusLink and RefSeq." Nucleic Acids Res **28**(1): 126-8.

Meyer, I. M. and R. Durbin (2002). "Comparative ab initio prediction of gene structures using pair HMMs." Bioinformatics **18**(10): 1309-18.

Modrek, B. and C. Lee (2002). "A genomic view of alternative splicing." Nat Genet **30**(1): 13-9.

Modrek, B. and C. J. Lee (2003). "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss." Nat Genet **34**(2): 177-80.

Modrek, B., A. Resch, et al. (2001). "Genome-wide detection of alternative splicing in expressed sequences of human genes." Nucleic Acids Res **29**(13): 2850-9.

Mott, R. (1997). "EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA." Comput Appl Biosci **13**(4): 477-8.

Parra, G., P. Agarwal, et al. (2003). "Comparative gene prediction in human and mouse." Genome Res **13**(1): 108-17.

Pruitt, K. D., K. S. Katz, et al. (2000). "Introducing RefSeq and LocusLink: curated human genome resources at the NCBI." Trends Genet **16**(1): 44-7.

Pruitt, K. D. and D. R. Maglott (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucleic Acids Res **29**(1): 137-40.

Pruitt, K. D., T. Tatusova, et al. (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **33**(Database issue): D501-4.

Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proc IEEE. **77**(2): 257-285.

Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." Proc. IEEE **77**(2): 257-286.

Reese, M. G., G. Hartzell, et al. (2000). "Genome annotation assessment in Drosophila melanogaster." Genome Res **10**(4): 483-501.

Reese, M. G., D. Kulp, et al. (2000). "Genie--gene finding in Drosophila melanogaster." Genome Res **10**(4): 529-38.

Rogic, S., B. F. Ouellette, et al. (2002). "Improving gene recognition accuracy by combining predictions from two gene-finding programs." Bioinformatics **18**(8): 1034-45.

Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-86.

Sachidanandam, R., D. Weissman, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-33.

Salamov, A. A. and V. V. Solovyev (2000). "Ab initio gene finding in Drosophila genomic DNA." Genome Res **10**(4): 516-22.

Schlueter, S. D., Q. Dong, et al. (2003). "GeneSeqer@PlantGDB: Gene structure prediction in plant genomes." Nucleic Acids Res **31**(13): 3597-600.

Schultz, J., T. Doerks, et al. (2000). "More than 1,000 putative new human signalling proteins revealed by EST data mining." Nat Genet **25**(2): 201-4.

Seki, M., M. Satou, et al. (2004). "RIKEN Arabidopsis full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions." J Exp Bot **55**(395): 213-23.

Snyder, E. E. and G. D. Stormo (1993). "Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks." Nucleic Acids Res **21**(3): 607-13.

Snyder, E. E. and G. D. Stormo (1995). "Identification of protein coding regions in genomic DNA." J Mol Biol **248**(1): 1-18.

Solovyev, V. and A. Salamov (1997). "The Gene-Finder computer tools for analysis of human and model organisms genome sequences." Proc Int Conf Intell Syst Mol Biol **5**: 294-302.

Solovyev, V. V. (2002). Finding genes by computer: probabilistic and discriminative approaches. Current Topics in Computational Biology. T. Jiang, T. Smith, Y. Xu and M. Zhang. Cambridge, MA, The MIT Press**:** 365-402.

Solovyev, V. V., A. A. Salamov, et al. (1994). "Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames." Nucleic Acids Res **22**(24): 5156-63.

Staden, R. (1984). "Computer methods to locate signals in nucleic acid sequences." Nucleic Acids Res **12**(1 Pt 2): 505-19.

Stanke, M., O. Schoffmann, et al. (2006). "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources." BMC Bioinformatics **7**: 62.

Stanke, M. and S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel." Bioinformatics **19 Suppl 2**: II215-II225.

Stein, L., P. Sternberg, et al. (2001). "WormBase: network access to the genome and biology of Caenorhabditis elegans." Nucleic Acids Res **29**(1): 82-6.

Stein, L. D., Z. Bao, et al. (2003). "The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics." PLoS Biol **1**(2): E45.

Stormo, G. D. (2000). "Gene-finding approaches for eukaryotes." Genome Res **10**(4): 394-7.

Stormo, G. D. and D. Haussler (1994). "Optimally parsing a sequence into different classes based on multiple types of evidence." Proc Int Conf Intell Syst Mol Biol **2**: 369-75.

Stormo, G. D., T. D. Schneider, et al. (1982). "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli." Nucleic Acids Res **10**(9): 2997-3011.

Strausberg, R. L., E. A. Feingold, et al. (2002). "Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences." Proc Natl Acad Sci U S A **99**(26): 16899-903.

Strausberg, R. L., E. A. Feingold, et al. (1999). "The mammalian gene collection." Science **286**(5439): 455-7.

Tatusova, T. A., I. Karsch-Mizrachi, et al. (1999). "Complete genomes in WWW Entrez: data representation and analysis." Bioinformatics **15**(7-8): 536-43.

The MGC Project Team (2004). "The Status, Quality, and Expansion of the NIH Full-Length cDNA Project: The Mammalian Gene Collection (MGC)." Genome Res **14**(10b): 2121-2127.

Uberbacher, E. C. and R. J. Mural (1991). "Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach." Proc Natl Acad Sci U S A **88**(24): 11261-5.

Usuka, J., W. Zhu, et al. (2000). "Optimal spliced alignment of homologous cDNA to a genomic DNA template." Bioinformatics **16**(3): 203-11.

Van Baren, M. J. and M. R. Brent (2005). "Iterative gene prediction and pseudogene removal improves genome annotation." Submitted.

Viterbi, A. J. (1967). "Error Bounds for convolutional codes and an asymptotically optimal decoding algorithm." IEEE Trans. Informat. Therory **IT-13**: 260-269.

Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Wei, C. and M. R. Brent (Submitted). "Integrating EST alignments and de novo gene prediction using TWINSCAN." Genome Res **15**.

Wei, C., P. Lamesch, et al. (2005). "Closing in on the *C. elegans* ORFeome by Cloning TWINSCAN predictions." Genome Res **15**: 577-582.

Wei, C., P. Lamesch, et al. (2005). "Closing in on the C. elegans ORFeome by cloning TWINSCAN predictions." Genome Res **15**(4): 577-82.

Wheelan, S. J., D. M. Church, et al. (2001). "Spidey: a tool for mRNA-to-genomic alignments." Genome Res **11**(11): 1952-7.

Wolfsberg, T. G. and D. Landsman (1997). "A comparison of expressed sequence tags (ESTs) to human genomic sequences." Nucleic Acids Res **25**(8): 1626-32.

Wu, J. Q., D. Shteynberg, et al. (2004). "Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing." Genome Res **14**(4): 665-71.

Xu, Y., J. R. Einstein, et al. (1994). "An improved system for exon recognition and gene modeling in human DNA sequences." Proc Int Conf Intell Syst Mol Biol **2**: 376-84.

Yeh, R. F., L. P. Lim, et al. (2001). "Computational inference of homologous gene structures in the human genome." Genome Res **11**(5): 803-16.

Zhang, M. and W. Gish (2006). "Improved spliced alignment from an information theoretic approach." Bioinformatics **22**(1): 13-20.

Zhang, M. Q. (2002). "Computational prediction of eukaryotic protein-coding genes." Nat Rev Genet **3**(9): 698-709.

Vita

Chaochun Wei

Birth Place: Zhejiang Province, China

Birth Date: 1973

Education:

D.Sc. in Computer Science (May 2006), Washington University in St. Louis

M.S. in Computer Science (May 2002), Washington University in St. Louis

M.E in Signal and Information Processing (July 1999), Beijing University

B.S. in Mathematics (July 1996), Beijing University, China

Publications:

1. "A new Spliced Alignment Algorithm Using Sequence Quality Values", Wei, C. and Brent, M. R. (2006) (In preparation).

2. "Using Expressed Sequence Tags to Improve Gene Structure Prediction", Wei, C. and Brent, M. R. (2006) (Submitted).

3. "PAIRAGON + N-SCAN: A Model-Based Gene Annotation Pipeline", Arumugam, M., Wei, C., Brown, R. H. and Brent, M. R. (2005) (Submitted).

4. "Closing in on the *C.elegans* ORFeome by Cloning TWINSCAN predictions", Wei, C., Lamesch, P., Arumugam M., Rosenberg, J., Hu, P., Vidal, M., and Brent, M. R. (2005) *Genome Research* 15:577-582.

5. "The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics", Stein, L. D., Z. Bao, et al. (2003). *PLoS Biol* 1(2): E45.

May 2006

Short Title: ESTs for Gene Prediction Chaochun Wei, D.Sc. 2006