Computational studies with ESTs: assembly, SNP detection, and applications in alternative splicing

Jianmin Wang
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/rtdin
Part of the Bioinformatics Commons

Recommended Citation
Computational studies with ESTs: assembly, SNP detection, and applications in alternative splicing

by

Jianmin Wang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Xiaoqiu Huang, Co-major Professor
Xun Gu, Co-major Professor
Eve Wurtele
Hui-Hsien Chou
David Fernández-Baca

Iowa State University
Ames, Iowa
2006

Copyright © Jianmin Wang, 2006. All rights reserved.
Graduate College
Iowa State University

This is to certify the dissertation of

Jianmin Wang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For Major Program
# TABLE OF CONTENTS

## ACKNOWLEDGEMENTS

## ABSTRACT

### CHAPTER 1 GENERAL INTRODUCTION

1. Introduction
2. Thesis organization
3. Literature review
   3.1. EST resources and tools
      3.1.1. The importance of EST sequences in biological research
      3.1.2. Major EST related databases
      3.1.3. EST clustering strategies
      3.1.4. EST assembly programs
      3.1.5. The significance of deep coverage assembly program
   3.2. EST applications in SNP discovery
      3.2.1. The feature of SNPs
      3.2.2. Major SNP databases
      3.2.3. SNP identification softwares
   3.3. Applications of ESTs in alternative splicing
      3.3.1. The biological roles of alternative splicing
      3.3.2. The prevalence of AS in metazoans
      3.3.3. Bioinformatics approaches for AS detection
      3.3.4. Analysis of AS variants
      3.2.5. Difficulties and challenge in AS study
4. References

### CHAPTER 2 AN ASSEMBLY PROGRAM FOR SEQUENCES WITH DEEP COVERAGE

Abstract
Introduction
Methods
Results
Discussion
Acknowledgements
References

### CHAPTER 3 A METHOD FOR FINDING SINGLE-NUCLEOTIDE POLYMORPHISMS WITH ALLELE FREQUENCIES IN SEQUENCES OF DEEP COVERAGE

Abstract
ACKNOWLEDGEMENTS

I thank my major professor Dr. Xiaoqiu Huang for instructions, guidances and financial support during my Ph D study. I also like to thank my co-major professor Dr. Xun Gu for the cooperation and help. I also like to thank all my present and former POS committee members: Dr. Aluru Srinivas, Dr. Eve S Wurtele, Dr. Hui-Hsien Chou, and Dr. David Fernandez-Baca for providing helpful suggestions in my research. In addition, I would like to thank Liang Ye, Zhixi Su and Haining Lin for helping in discussions and suggestions in the research.

Finally, I would like to thank my family and friends for their help and support, especially my wife Ying Ye and my lovely daughter Irene Wang.
ABSTRACT

EST sequences are important in functional genomics studies. To better use available EST resources, clustering and assembling are crucial techniques. For EST sequences with deep coverage, no current assembly program can handle them well. We describe a deep assembly program named DA. The program keeps the number of differences in each contig alignment under control by making corrections to differences that are likely due to sequencing errors. Experimental results on the 115 clusters from the UniGene database show that DA can handle data sets of deep coverage efficiently. A comparison of the DA consensus sequences with the finished human and mouse genomes indicates that the consensus sequences are of acceptable quality.

EST sequences can be used in SNP discovery. We describe a computational method for finding common SNPs with allele frequencies in single-pass sequences of deep coverage. The method enhances a widely used program named PolyBayes in several aspects. We present results from our method and PolyBayes on eighteen data sets of human expressed sequence tags (ESTs) with deep coverage. The results indicate that our method used almost all single-pass sequences in computation of the allele frequencies of SNPs.

EST sequences can also be used to study alternative splicing (AS), which is the most common post transcription event in metazoans. We first developed a pipeline to identify AS forms by comparing alignments between expressed sequences and genomic sequences. Then we studied the relationship between AS and gene duplication. We observed that duplicate genes have fewer AS forms than single-copy genes; we also found that the loss of alternative splicing in duplicate genes may occur shortly after the gene duplication. Further analysis of
the alternative splicing distribution in human duplicate pairs showed the asymmetric evolution of alternative splicing after gene duplications. We also compared AS among six species. We found significant differences on both AS rates and splice forms per gene among the studied species by detailed and categorized studies. The difference in AS rate between rice and Arabidopsis is significant enough to lead to a difference in protein diversity between those two species.
CHAPTER 1 GENERAL INTRODUCTION

1. Introduction

An EST (Expressed Sequence Tag) is a single pass read from the 5’ or 3’ end of a random cDNA clone with a length from 300 to 500 bases. With the improvement of sequencing techniques, high volume EST data are generated at really low costs. EST sequences are from gene coding regions of genomes, they provide an inexpensive approach to the studies of gene structure, alternative splicing, gene expression and single nucleotide polymorphisms (SNPs). Even for organisms with genomes being sequenced, the EST data still provide invaluable information in functional genomics.

To provide an easy to use and unified access to EST resources for different purposes, storing, clustering and assembling ESTs systematically are important. The dbEST database is the major repository for depositing and querying EST sequences. It does not have any processing so it is highly redundant and contains different kinds of errors. To group sequences from the same gene together and to reduce the redundancy and errors in EST data sources, several EST related projects were carried out. Those projects include TIGR Gene Index (Quackenbush et al. 2000), NCBI UniGene (Pontius 2003) database and STACKdb (Christoffels et al. 2001). All these projects use different clustering techniques and some of them also include an assembly program to generate consensus sequences.

EST clustering and assembling have the central roles in reducing the redundancy of EST sequences and generate high quality consensus sequences. EST clustering is finding and grouping EST sequences from the same gene. There are several clustering tools. UniGene, TIGR Gene Index and SANBI STACK use non-parallel clustering programs. PaCE
(Kalyanaraman et al. 2003) is a parallel clustering program and Ulcluster (Trivedi et al. 2003) can run in both parallel and serial modes. All EST clustering methods are based on the fact that EST sequences from the same gene are highly similar. Among EST assembly programs, Phrap (http://www.phrap.org/), TIGR assembler (Sutton et al. 1995) and CAP3 (Huang and Madan 1999) are the most commonly used. They all use similar approaches with three steps: overlap computation, contig construction and multiple alignment output. By using EST assembly programs, accurate consensus can be derived. EST assembly programs can also be used as clustering tools, and sequences in the same contig are grouped together.

SNPs are useful in different association studies and population genetics. As a genetic markers SNPs can reach the highest resolution with one SNP in 300 to 1200 bases, which is extreme useful in multiple genes related association studies. Each SNP most time has two different alleles and both are dominant, so genetic drift will change allele frequencies in different populations and this can be used to distinguish different populations. Compared with other polymorphisms SNPs are more stable, so SNPs are more reliable in genetic studies. SNPs can appear at any position of genome such as intron, exon, intergenic region and regulatory region. So SNPs can change phenotypes and sometimes are related to diseases directly or indirectly. Genotyping of SNPs is easy with the microarray technique; Affymatrix has developed chips which can genotype 100,000 SNPs on a single chip. Those features make SNP one of the most important genetic markers in lots of studies.

SNP detection using genomic sequences requires lots of sequencing efforts, on the other hand SNP detection using expressed sequences can utilize the large amount of available data and can help identify some important disease related SNPs especially non-synonymous SNPs. For genomic regions with high EST coverage, frequencies of the alleles can be
calculated to provide additional information. Currently there are two different types of SNP detection programs, detection of SNPs from cloned samples and detection of SNPs from diploid samples. Most programs can’t handle EST sequences with deep coverage well due to the error-prone nature of EST sequences.

In animals alternative splicing (AS) is the most common post-transcriptional process which greatly increases the complexity of transcriptome and the number of proteins (Maniatis and Tanis 2002). Different studies on human reported the AS rate from 30% to 70% and all the studies showed the prevalence of AS in human (Mironov et al. 1999; Brett et al. 2000; Croft et al. 2000; International Human Genome Sequencing Consortium 2001; Kan et al. 2001; Modrek et al. 2001). Also studies on functional analysis and tissue specificity (Xu et al. 2002; Xie et al. 2002) of AS were carried out. The comparisons of AS structure and conservation (Kriventseva et al. 2003; Valenzuela et al. 2004) among species were used in evolutionary studies.

Even though AS is common in animals, the differences among species were not well addressed. The relationship between AS and the complexity of organisms was studied by two groups of researchers with controversial results. Brett et al. (2002) used 650 mRNA sequences and 100,000 ESTs from each of the seven species to compute AS rates and splice forms per gene. The EST coverage for each mRNA ranges from 3 to 20. In their study, no significant differences of AS rates and the numbers of splice forms per gene between human and other animals was observed. But significant differences between mammals and invertebrates were discovered by Harrington et al. (2004), they computed the average number of splice forms per gene by estimating the total number of proteins.
Another interesting question is what are the AS rates in plants? A recent study (Iida et al. 2004) on Arabidopsis showed that 11.6% of the genes were alternative spliced. Another study (Kikuchi et al. 2003) showed that 13.1% of rice genes are alternative spliced. So plant might have higher AS rates than the predicted 5% (Kazan 2003), but the lacks of genomes, full length cDNAs and ESTs impose difficulties on AS studies of other plant species.

In the present study, our goals are: 1) Develop an EST assembly program that can handle sequences with deep coverage; 2) Develop a SNP detection program that can deal with sequences with deep coverage, and the program can also identify sequences from different paralogous regions and report SNP allele frequencies; 3) Develop a pipeline to identify different splice forms from the alignments between genomes and expressed sequences; 4) Comparison studies of AS using the pipeline, including AS comparisons among six species and comparison between AS and gene duplication.

2. Thesis organization

The thesis consists of six chapters. All those chapters focus on the different aspects of ESTs, the central biological data type through this study.

In Chapter 1, background knowledge of EST is introduced. It includes: the importance of EST sequences, computational techniques for EST studies and EST applications in biological studies. Also the gaps between previous works and biological requirements are revealed to verify the purpose of our research.

We describe an EST assembly program for sequences with deep coverage in Chapter 2. The program (DA) couples the overlap computation and the contig construction processes, which are separate steps in previous approaches. DA uses a high quality alignment
immediately to either build a new contig or add an un-grouped sequence into an existing contig. The program utilizes a data structure called profiles to represent each contig, which can help to save space and simplify the alignment involving contigs. The time efficiency and accuracy of the program were evaluated by 114 datasets downloaded from UniGene cluster and one dataset from TIGR. DA computes only a linear number of alignments relative to the number of input sequences and it generates accurate consensus sequences. The project was proposed by John Quackenbush, the algorithm and program were developed by Jianmin Wang under direction and supervision of Xiaoqiu Huang. Geo Pertea tested the program.

In Chapter 3, we present a computer program, PolyFreq, which identifies SNPs by comparing alignments between sequences from different individuals and anchor genome sequences. The program uses informative sites and similarities to distinguish the true corresponding genome location from possible paralogous locations for each individual sequence. Also the program does not compute multiple alignments, it uses a profile representation to help identify SNPs and avoid possible hassles of building multiple alignments. The algorithm and program were developed by Jianmin Wang under direction and supervision of Xiaoqiu Huang.

Then the applications of EST in alternative splicing (AS) are presented in Chapter 4 and 5. We first describe a pipeline to identify alternative splice forms by comparing the alignments between genomes and expressed sequences. The pipeline was developed by Jianmin Wang under direction of Xiaoqiu Huang. Then we studied the evolutionary trend of AS after gene duplication. We observed that duplicate genes have fewer AS forms than single-copy genes; we also find that the loss of AS in duplicate genes may occur shortly after gene duplication. These results support the subfunctionization model of AS in the early stage.
after gene duplication. Further analysis of the AS distribution in human duplicate pairs shows
the asymmetric evolution of AS after gene duplications. Jianmin Wang identified splice
forms using the pipeline and identified duplicated genes. Zhixi Su did the statistical and
evolutional analysis. This study is under the direction of Jun Yu, Xiaoqiu Huang, and Xun
Gu. We also compared AS among six species: human, mouse, fruit fly (*Drosophila*),
nematode (*C. elegans*), rice, and mustard (*Arabidopsis*). We find significant differences on
both AS rates and splice forms per gene among the studied species by detailed and
categorized studies. We discover that AS rate in Arabidopsis is similar to previous study, but
rice has a much higher AS rates than previous study. The difference of AS between rice and
Arabidopsis is significant, so the diversity among plants could be tremendous. This study
was done by Jianmin Wang under direction and supervision of Xiaoqiu Huang. Liang Ye
helped in data gathering and results verification. Jianjun Chen gave useful suggestions in
present the results. Xun Gu helped with statistical methods used in the study.

Chapter 6 is the general conclusion for all studies in this dissertation. It summarizes the
results of several studies and points out possible future research directions.

### 3. Literature Review

This dissertation concentrates on studies of EST sequences; which include a
computational tool to assemble EST sequences with deep coverage, a program to detect
SNPs, and applications in AS using EST sequences. The first part of the review focuses on
EST resources, clustering strategies and assembly programs; the second part of the review
introduces SNPs and SNP detection programs; the last part of the review contains previous
experimental and bioinformatics studies of AS. Each part of the review also points out existing problems that have been not well addressed.

3.1. EST resources and tools

3.1.1 The importance of EST sequences in biological research

EST (Expressed Sequence Tag) is a single pass read from the 5’ or 3’ end of a cDNA clone. A basic method for fast EST sequencing was first described by Adams et al. (1993). Methods for constructing cDNA libraries and strategies of doing normalization and library subtraction to increase the representation of lowly transcribed mRNA were described in detail by Baldo et al. (1996). With the improvement of sequencing techniques, high volume EST data are generated at really low costs by different groups. With the genome sequences available for many species, EST sequences play a central role in biological researches to identify genes in finished genomes. EST sequences are from gene coding regions of genome, they can be used in different research areas (Adams et al. 1991).

ESTs can be used for gene prediction, gene structure study and genomic mapping (Kulp et al. 1996; Jiang and Jacob 1998; Brendel et al. 2004). Most gene prediction programs require known gene structures to train the parameters. Alignments between EST sequences and genomic sequences provide exact splice sites and some EST sequences have poly-A tails that cover terminal exons. Also as genetic markers, ESTs can be mapped to genomic region based alignments to help build more detailed genetic maps.

ESTs can be used to characterize alternative splicing, which is a major source of multiple gene products from the same gene (Modrek et al. 2001; Xie et al. 2002). Computational identification of AS relies on alignments of genomic sequences and expressed
sequences. Alternative splice forms can be identified and categorized by comparison of splicing patterns between different expressed sequences.

ESTs can be used to study gene expression and regulation in different tissues, developmental stages and environments (Ewing and Claverie 2000; Vasmatis et al. 1998). The number of EST sequences for the same gene from libraries under different conditions can serve as the proxy of expression level to do expression profile study. Furthermore partial unique EST sequences can be used in SAGE (Serial Analysis of Gene Expression) to study gene expression (Velculescu et al. 1995; Velculescu et al. 1997).

ESTs can be used to identify SNPs (Single Nucleotide Polymorphisms), especially for SNPs that are related to non-synonymous changes (Picoult-Newberg et al. 1999). For regions with deep EST coverage it's also possible to compute allele frequencies. By aligning EST sequences to corresponding genomic sequences it's easy to identify potential SNPs, but statistical methods are required to distinguish sequencing errors and SNPs. Also experimental verifications of SNPs are important and are also efficient by using microarray technology.

3.1.2. Major EST related databases

dbEST is the major repository for EST sequences, it contains EST sequences for dozens of species. Currently there are 32,889,225 EST sequences in dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html, Jan. 20, 2006). The number of ESTs keeps on increasing rapidly each year. Human, mouse and Xenopus tropicalis have over 1 million EST sequences among all species. There is no sequence processing in dbEST, so the redundancy is high and also the sequences may contain different types of errors.
To give an easy to use and unified view of EST data resources, several EST related projects were carried out. Most projects cluster EST sequences into groups such that in every group the EST sequences come from the same gene or gene family. In each cluster a consensus sequence is generated by an assembly program. There are three major databases that cluster similar EST sequences together using different strategies. TIGR Gene Index (TGI) (Quackenbush et al. 2000) is a project that tries to group all ESTs and known gene sequences from GenBank concerning one gene into a single index class. Each splicing variant is considered as one gene. Each gene index has one tentative consensus (TC) sequence and is annotated. Currently TGI contains 88 species (31 animals, 33 plants, 9 fungi and 15 protists, March 2006). As a comprehensive data resource, TGI uses many computer programs for different purposes and provides annotations at different levels. The NCBI UniGene (Pontius 2003) cluster is another important resource of EST clustering, it uses several kinds of GenBank sequences and has no consensus for each single cluster since different splicing variants and possibly different genes may be in the same cluster. Currently (Mar, 2006) UniGene contains 71 species with over 1 million clusters. UniGene is also a comprehensive database with lots of cross links to other databases and detailed annotations. STACKdb (Christoffels et al. 2001) is another database that uses stackPACK to mask, cluster and assemble (by Phrap) EST sequences. It can also detect alternative splicing.

3.1.3. EST clustering strategies

The clustering methods used in the previous three databases are all based on similarities between sequences. They are all sequential methods and they have different criteria to build clusters. UniGene clusters are non-incremental, so they need to be rebuilt for each new
version. There are two different strategies to build clusters in UniGene, transcript based and genome based. For transcript based clustering, megaBLAST (Zhang et al. 2000) is used to compare mRNA sequences and compare mRNA with EST sequences. It also requires a polyadenylation signal or tail in each cluster. Genome based clustering is used for species with genome sequences available. Gene boundaries and corresponding transcripts are identified by using several types of evidences. First, gene annotations including RefSeqs and transcripts with predicted proteins are used. Second, transcribed sequences that can be well aligned to genomes with splice site consensus are used to identify further splice boundaries. Third, unspliced transcripts and sequences that can’t be aligned to genomes are considered too. TIGR Gene Index uses mgBLAST, a modified version of megaBLAST (Zhang et al. 2000), to align each pair of sequences. The alignments are used to build clusters. Also it uses Paracel Transcript Assembler, a modified version of CAP3 (Huang and Madan 1999), as the assembly program to generate TCs. STACKdb has a loose requirement to merge clusters, but is stringent on assembly. The stackPACK uses a non-alignment method called d2_cluster (Hide et al. 1994; Burke et al. 1999) to do the clustering. The d2_cluster method starts with each sequence in a cluster and keeps on merging clusters according to the distance between each pair of sequences. The criterion for merging clusters is based on the existence of percent identity in a window between two sequences. The method can be described as minimal linkage clustering and is relatively loose to merge two clusters. Essentially all sequence assembly programs can be used to cluster EST sequences too.

There are also programs that use parallel computing techniques to accelerate the clustering process. One such program is PaCE (Parallel Clustering of ESTs) developed by Kalyanaraman et al. (2003). The program uses General Suffix Tree (GST) to find similar
sequence pairs and distributes GST at each node. The overall hardware structure consists of one master node and multiple slave nodes. The GST is built on each slave node and promising pairs are generated by slave nodes too. The master node determines batches of sequence pairs to be aligned and slave nodes compute alignments. Then according to the alignment, the master node determines whether two sequences should be put into the same cluster. The method has improvements on both running time and space requirement by building GST and computing alignments in parallel. Another program is Ulcluster (Trivedi et al. 2003), which has both sequential and parallel versions. The program compares each input sequence with the primary sequence of every cluster. If the similarity is high enough, the sequence is put into the corresponding cluster. Otherwise a new cluster with only the input sequence is created. The method of finding similar pairs quickly is by requiring \( n \) matches in an \( m \) word window; a hashing table and a map containing those values are also used. The program has two different parallel methods. One is parallelization on cluster space. All clusters are evenly distributed on each node; each incoming sequence is processed on every node. When matches are found, the alignment information is communicated among nodes and the best one is used to add the incoming sequence into the corresponding cluster. Another is parallelization on input space. The input sequences are evenly distributed on each node. This is similar to the sequential version but with a smaller dataset.

3.1.4. EST assembly programs

The assembly program is a key component in the process to reduce EST redundancy and generate long accurate consensus sequences because ESTs are error prone and are relatively short. Currently there are three major assembly programs for EST assembly, Phrap
(http://www.phrap.org/), CAP3 (Huang and Madan 1999) and TIGR Assembler (Sutton et al. 1995). All these programs use similar strategies. A fast alignment algorithm is first used to find significant pairs, then detailed alignments are computed and sorted according to different criteria. Then a greedy method is used to build contigs by choosing the current best alignment from all unused alignments. At last for each contig a multiple alignment is constructed and a consensus sequence is generated. A comparison of the three assembly programs (Liang et al. 2000) showed that CAP3 generated the best results on over 118k rat EST sequences.

3.1.5. The significance of deep coverage assembly program

EST sequences are error prone because of their sources and single pass nature. The average error rate of EST sequences is around 2 percent. An obvious problem with EST sequences is that they are not evenly distributed; some highly expressed genes have many more ESTs. Also lots of genes have multiple copies, and paralogous genes or pseudogenes can increase the coverage too. In the UniGene database, there are 1974 clusters with more than 1000 sequences in January 2006 release.

Groups of very deep coverage impose two computational challenges. Firstly, the number of overlaps between sequences of deep coverage is very huge. It takes the existing programs huge amounts of time and space to compute and save overlaps. Second, the deep coverage and sequencing errors in EST sequences are problematic to the methods used in the existing programs for generating consensus sequences. For example, the CAP3 program produces a multiple sequence alignment for each contig and generates a consensus sequence from the alignment for the contig. A multiple alignment of EST sequences with deep coverage is hard
to build because the percent identity of the alignment is very low. Since each position of the alignment is covered by thousands of sequences, there is likely to be a gap column between any two match columns in the alignment. These gap columns are due to insertion sequencing errors in sequences. Thus, nearly half of the columns in the alignment are gap columns and hence the percent identity of the alignment is below 60%. If the percent identity of a DNA alignment is below 60%, the alignment can not be accurately computed by a standard alignment algorithm.

Currently most EST projects use the three programs mentioned in the previous section as the assembly program to produce none-redundant consensus sequences. These programs were developed to assemble sequences with relative low coverage. For a big EST cluster with deep coverage, all these programs need too much time to finish. And even the sequences are assembled; the high error rate of EST sequences makes the assembly results hard to do further analysis given the fact that potentially every position will have a gap. So an assembly program, which can handle deep coverages and high error rates, is important in analysis of EST sequences.

We develop a deep assembly program named DA for assembling sequences of deep coverage. Instead of computing and evaluating a huge number of possible overlaps among sequences, DA repeatedly computes best overlaps and immediately uses them in construction of contigs. Each contig is represented by a base count profile and a quality profile. The DA program keeps the number of differences in each contig under control by making corrections to differences that are likely due to sequencing errors. For each contig, DA reports a multiple sequence alignment, a profile, a consensus sequence, and the positions of popular differences in the contig. Experimental results on the clusters from NCBI UniGene and one large cluster
from TIGR Gene Index show that DA can handle data sets of deep coverage efficiently. Comparisons of the DA consensus sequences with the finished human genome indicate that the consensus sequences are of acceptable quality. The DA program can also generate splice variants.

3.2. EST applications in SNP discovery

3.2.1. The features of SNPs

A single nucleotide polymorphism (SNP) is a single base variation in genomic sequences. This kind of variations can be used as a genetic marker like other polymorphisms in association studies. Compared with other genetic markers, SNP has some features that are more suitable for complicated multiple gene related association studies.

SNP has the highest density and is the most abundant among all genetic markers. It was estimated that there is 1 SNP for every 300 to 1200 bases in human genome (Lander et al. 2001) and SNPs make up 80% of known polymorphisms. With the genome resequencing projects, it will not be a surprise to see even more SNPs in genomes. The high density of SNPs on genomes can help positional cloning for gene mapping studies. Closely located SNPs can be used to study linkage disequilibrium (LD). The high density of SNPs also makes them very useful in gene discovery and multiple gene related diseases studies such as diabetes, cardiovascular diseases and most cancers. Those studies require makers with enough density to cover the whole genome.

SNP can occur at any position of a genome such as introns, exons, coding regions and intro-generic regions. This makes SNP an important marker in both population genetics studies and molecular biology studies. SNPs appearing in exons may be non-synonymous
changes and thus can change protein sequences. So some SNPs are related to functional changes and furthermore to phenotype changes.

A SNP mostly has two alleles and both are dominant. The small number of alleles is due to the low base mutation rate, which is 1E-9 to 5E-9 for mammals (Li et al. 1981; Martinez-Arias et al. 2001). The low base mutation rate makes it possible to test the relationship between phenotypes and functional differences because high mutation rate may compromise association studies (Marshall et al. 1993). Also transitions (purine to purine and pyrimidine to pyrimidie) are more common than transversions (purine to pyrimidine), and a study shows that transition to transversion ratio is 1.7 in human (Picoult-Newberg et al. 1999).

SNPs are easy to genotype. With neighbor sequences known, it’s easy to design experiments to genotype any SNP. The major methods include hybridization, primer extension, ligation and invasive cleavage. Each of the methods can be extended to array based analysis with different support materials and detection methods. The microarray genotyping is using large scale hybridizations. It was developed by Whitehead Institute and Affymetrix together (Wang et al. 1998). Also the new microarray product, GeneChip HuSNP, can genotype 100,000 SNPs on one single chip.

3.2.2. Major SNP databases

There are two major SNP databases, dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/index.html) and HapMap. dbSNP is the major repository for all species and currently it has 35 species (release 125) with total 40 million SNPs. There are six species which have over 1 million SNPs. For most species only a limited number of SNPs have been
validated (18 million are validated), also only fewer of the SNPs have allele frequencies information (1.3 millions). Currently, dbSNP also has the haplotype information from papers. HapMap (The International HapMap Consortium 2005) is an international effort to identify and catalog genetic differences of human beings. 270 individuals were sampled from Nigeria, Japan, China and US. SNPs serve as proxies in HapMap project to help identify different haplotypes. In this project, the target SNPs should have a minor allele frequency greater than 5%. There are over 1 million SNPs that passed the quality control requirements in HapMap. These two data resources can be used as references but they do not have a complete list of all SNPs.

3.2.3. SNP identification softwares

Identification of SNPs using genomic sequences from different individuals needs lots of resequencing efforts and may be cost inefficient. The ability to find SNPs from heterozygotes helps to identify more SNPs. Available EST sequences can also help to identify SNPs with low costs and to estimate allele frequencies for genomic regions with high coverage. The high error rates of ESTs and rare RNA editing may cause extra false positive SNPs. To mine SNPs from resequencing projects or existing EST sequences, several softwares were developed to identify potential SNPs by comparing sequences from individuals with known reference genome sequences. There are two major types of SNP detection softwares: 1) detections from cloned samples, which focus on distinguishing sequencing errors from real variants; 2) detections from diploid samples, which focus on distinguishing sequencing errors from heterozygous sequences. All those programs compute a multiple alignment first, and do different analysis using either quality values or chromographs directly.
There are several programs designed to identify SNPs from EST data. Picoult-Newberg et al. (1999) described their method to identify and verify SNPs. They assembled EST sequences using Phrap and the multiple alignments of assembled results were viewed by Consed. Several filters are used to deal with sequencing and alignment errors, such as mismatches from low-quality regions, sequence mismatches by base substitution or indel, and base calling qualities. They also designed experiments to verify 20% of the identified SNPs. Their method does not have a reference sequence and involves assembly which may cause errors. PolyBayes (Marth et al. 1999) is another program that deals with EST sequences or genomic sequences. PolyBayes first uses CROSS_MATCH (in Phrap package) to align EST sequences with the anchor genomic sequence. Then based on these pairwise alignments, a multiple alignment is built by incorporating gaps. Then the program identifies paralogue by determining whether the number of mismatches observed in an EST is consistent with polymorphism or paralogue. The posterior distribution of native EST (non-paralogue) under the uniform prior is computed and compared with a cutoff value to determine whether an EST belongs to the genomic anchor or not. Then for each column in the multiple alignments the program identifies a SNP by computing the likelihood of being heterogeneous.

There are also some programs that focus on identification of SNPs from genomic sequences. PolyPhred (Stephens et al. 2006) is such a program that uses Phred, Phrap and Consed to help in base calling, alignment, viewing and editing. The program first identifies a “consensus base” at each position for each multiple alignment. Then the program computes measures of strength to be homozygous (same or different from consensus) and heterozygous
at each position for each sequence. Further statistical methods are applied to compute the
probability of each genotype by Bayes statistics under the Hady-Weinberg prior.

novoSNP (Weckx et al. 2005) is another program that identifies SNPs and INDELs
from diploid samples with its own graphical interface. The program uses BLAST to align
sequences to a reference sequence and the program works on the trace file directly. The
program identifies SNPs by compute a score using a cumulative scoring scheme which is the
sum of three different subscores. Those subscores are used to capture the major features of a
SNP in each column. INDELs are treated separately by their typical patterns such as frame-
shift.

SNPdetector (Zhang et al, 2005) is another program to identify SNPs and INDELs from
genomic sequences. It uses Phred to do base calling and SIM to align each sequence to the
template sequence. Then the program identifies high quality differences in each sequence
using the neighborhood quality standard. At last the program identifies heterozygous
genotypes by computing the ratio of primary and secondary peaks in the trace file.

Even though there are several SNP detection programs available, they are not designed
to mine SNPs from EST sequences with deep coverage due to the error prone nature. Also
they do not estimate allele frequencies due to the limited number of sequences. We
developed a program, PolyFreq, which utilizes sequences with deep coverage to estimate
allele frequencies and builds profiles instead of multiple alignments from pair-wise
alignments. Also the program assigns each EST sequence to one paralogous genomic
sequence using informative sites such that false positive rates can be reduced.
3.3. Applications of ESTs in alternative splicing

3.3.1. The biological roles of alternative splicing

Alternative splicing (AS) is one of the most common post-transcriptional processes in animals and it can generate multiple transcript isoforms from one coding region. Sambrook (1977) found the adenovirus mosaic gene structures, and the concepts of intron and exon were defined by Gilbert (1978), who proposed different combinations of exons may produce new transcripts. At that time, AS was regarded as an exception. But after decades of study, people realized AS is a complicated and common biological process involving lots of proteins and small nucleotides. There are over 200 different types of post-transcriptional modification, and AS is one of the most common events and it plays a critical role in lots of biological processes. Since AS can change the sequences of protein products, it can also change the functions of gene products. Therefore AS can increase the diversity and complexity of proteomics. One example is the sex-determination system in *Drosophila*. Three genes (sex lethal:sxl, transformer:tra, and doublesex:dsx) have different AS patterns that lead to the development of female or male (Baker 1889; Boggs et al. 1987). Recent examples showed that AS can play an important role in apoptosis (Lopez 1998; Boise et al. 1993), acoustic tuning in the ear (Black 1998; Ramanathan et al. 1999; Graveley 2001), cancer development (Carstens et al. 2002) and other biological processes. A study by sampling 50 alternative spliced genes showed that 75% of them are related to signalings and regulations, and most of them are from immune and nerve systems (Modrek 2001). And an astonishing observation of fruit fly nerve system revealed that the combination of multiple cassettes of exons can generate up to 40,000 splice forms for a single gene (Graveley 2001).
3.3.2. The prevalence of AS in animals

After the whole human genome was sequenced, it was found that the estimated number of human genes is lower than expected. To reach the complexity of human organism, the genes should be regulated with multiple mechanisms. AS is one such mechanism that can increase the effective number of gene products. A study on eukaryotes AS by mapping ESTs onto mRNA sequences (Brett et al. 2002) showed that the portion of alternative spliced genes in eukaryotes ranges from 5% (plant) to 43% (human). Mironov et al. (1999) found 133 alternative spliced genes among 392 known genes by aligning ESTs with genomic sequences. Croft et al. (2000) found 582 alternative spliced genes by searching annotated intron sequences from GenBank against EST sequences. 145 alternative spliced genes were reported by the International Human Genome Sequencing Consortium from analysis of chromosome 22 by aligning ESTs to genomic sequences. Modrek et al. (2001) identified 6201 alternative splice forms in 2272 genes by the multiple alignment of EST, mRNA and genomic sequences; they also imposed stringent criteria for AS detection (high similarity, splice site detection etc.). These studies on human AS showed that the alternative spliced genes could be from 35% to 59%. With the increasing coverage of ESTs to the mRNA sequences, more AS transcripts are expected. Kan et al (2001) reported that given enough EST sequences almost all genes have alternative splice forms. They also claimed that only 17-28% genes have alternative splice forms which has frequency of 5% or higher. A recent study (Gupta et al. 2004) based on EST consensus sequences showed that 45% of EST clusters (total 33,270) are alternatively spliced.

While most of the studies focused on human AS, a recent research (Iida et al. 2004) on Arabidopsis showed that 11% of the genes (1764 over 15214) are alternatively spliced and
the study also showed that AS profiles change with the change of environment and development stages. Also another study (Kikuchi et al. 2003) showed that 13.1% of rice genes are alternative spliced using 28,000 cDNA clones. Those studies still show plants have relatively low AS rates and the diversity of AS is not clear.

3.3.3. Bioinformatics approaches for AS detection

Most bioinformatics studies of AS rely on the differences among EST sequences and mRNA sequences from the same gene to detect AS. And each possible splice form will be checked against further criteria. Some researchers utilize the finishing human genome draft to identify splicing patterns and to derive new AS transcripts (Kan 2002). All these methods are based on computation, and only a few studies did experimental verification. Brett et al (2000) verified the result of 20 predicted splicing genes by experiments, they found 4 of the genes did not show the alternative splice forms predicted by the computational analysis and 16 genes did have AS forms. Hu et al (2001) reported a method of the large scale verification and detection of AS events. They first predicted AS variants using microarray data on 10 different mouse tissues. Then they verified top 50% by RT-PCR and compared the result with EST clustering results. Johnson et al. (2003) used exon junction microarrays to verify alternative splicing for more than 10,000 human genes. They reported at least 74% multi-exon genes are alternative spliced.

3.3.4. Analysis of AS variants

Several groups of researchers studied the conservation, the tissue distribution and other aspects of AS. The AS comparisons between human and mouse, human and rat (Modrek and Lee 2003) showed that exons in the common forms (non-spliced form) are highly conserved,
but the exons are only included in AS forms are less conserved. So the AS is related with evolutionary changes. Valenzuela et al. (2004) studied the AS conservation in the protein function modulation among human, mouse, rat and fruit fly. And they found high degree of conservation between species. Xie et al. (2002) studied the tissue distribution of AS using EST tissue information. In 10 most AS abundant tissues of their study, only few of AS forms are absolutely tissue specific and exon skipping is not the most common form. Kriventseva et al. (2003) studied the structural information of AS, they found that most of AS forms insert or delete a whole protein domain instead of disrupting the domains. So they believe that positive selection works on AS.

The relationship between AS and the complexity of organisms was studied by two groups with controversial results. The study (Brett et al. 2002) by comparing percentage of spliced genes showed that the differences between human and other species are not significant. They used 650 mRNA sequences and 100,000 EST sequences from each of seven different species. Also they showed that the difference of number of splice forms per gene are not significant between human and mouse. But another study (Harrington et al. 2004) predicted number of spliced forms per gene by calculating the total number of genes, and the results for four species showed significant differences.

3.3.5. Difficulties and challenge in AS study

As one of the most important biological processes that involves protein function and evolution, AS has been studied for decades. But the complete view of AS at genomic and species level is far from finished. The false negative and false positive effects of computation prediction and experimental factors of AS (Modrek and Lee 2002) create lots of
complexities on the problem. Also most EST sequence data are base on cell lines, which in fact introduces another problem. The lack of large scale experimental verification methods makes the problem even harder. Even we try to ignore the error rate of prediction, we don’t have enough data to do the prediction and the data are not evenly distributed either. All these problems require more sophisticated methods to detect/predict AS and more feasible experimental methods to verify those detections/predictions.

In summary, AS is one important biological process that has dramatic impacts on protein functions and diversities. The high level view of AS may help us to understand and answer lots of interesting biological questions. But the incomplete data, the inaccurate computational methods and infeasible large scale experiment methods demand better methods to do the analysis.

4. References


CHAPTER 2 AN ASSEMBLY PROGRAM FOR SEQUENCES OF DEEP COVERAGE

Jianmin Wang, Geo Pertea, John Quackenbush, and Xiaoqiu Huang

ABSTRACT

Motivation: There are over 18 million EST sequences in dbEST. Clustering and assembly of EST sequences is a standard step in utilization of this resource. However, existing assembly programs can not handle EST sequences of deep coverage, where the depth of coverage is in thousands.

Results: We describe a deep assembly program named DA. Instead of computing and evaluating a huge number of possible overlaps between sequences, DA repeatedly compute best overlaps and immediately use them in construction of contigs. Each contig is represented by a multiple sequence alignment. The DA program keeps the number of differences in each contig alignment under control by making corrections to differences that are likely due to sequencing errors. Experimental results on the seven largest clusters from the UniGene database show that DA can handle data sets of deep coverage efficiently. A comparison of the DA consensus sequences with the finished human genome indicates that the consensus sequences are of acceptable quality. The DA program can also generate splice variants. Similarly good results were obtained for several TIGR tentative consensus sequences (TCs). We expect that DA will be a useful tool in analysis of EST sequences.

Availability: The source code of DA is freely available for academic use at http://bioinformatics.iastate.edu/aat/sas.html
Introduction

There are over 18 million EST (Expressed Sequence Tag) sequences in dbEST (Boguski et al. 1993). To make this data resource useful and easily accessible to the scientific community, a number of EST databases have been constructed: NCBI UniGene (Wheeler et al. 2003), TIGR Gene Index (Quackenbush et al. 2000), and STACKdb (Christoffels et al. 2001). In those database projects, EST and mRNA sequences are clustered into groups such that sequences in a group come from the same gene. Sequences in a group are usually assembled by a sequence assembly program into longer consensus sequences with fewer sequencing errors. Small groups of sequences can be easily handled by existing sequence assembly programs (Staden 1980; Peltola et al., 1984; Huang 1992; Green 1995; Kececioglu and Myers 1995; Sutton et al. 1995; Huang and Madan 1999; Kim and Segre 1999; Chen and Skiena 2000; Pevzner et al. 2001). However, groups of sequences of coverage depths in thousands can not be handled by existing sequence assembly programs (Pertea et al. 2003).

Groups of very deep coverage impose two computational challenges. Firstly, the number of overlaps between sequences of deep coverage is very huge. It takes the existing programs huge amounts of time and space to compute and save overlaps. Second, the deep coverage and sequencing errors in EST sequences are problematic to the methods used in the existing programs for generating consensus sequences. For example, the CAP3 program produces a multiple sequence alignment for each contig and generates a consensus sequence from the alignment for the contig. A multiple alignment of EST sequences with deep
coverage is hard to build because the percent identity of the alignment is very low. Since each position of the alignment is covered by thousands of sequences, there is likely to be a gap column between any two match columns in the alignment. These gap columns are due to insertion sequencing errors in sequences. Thus, nearly half of the columns in the alignment are gap columns and hence the percent identity of the alignment is below 60%. If the percent identity of a DNA alignment is below 60%, the alignment can not be accurately computed by a standard alignment algorithm.

Here we describe a deep assembly program named DA for assembling sequences of deep coverage. Instead of computing and evaluating a huge number of possible overlaps between sequences, DA repeatedly compute best overlaps and immediately use them in construction of contigs. Each contig is represented by a multiple sequence alignment. The DA program keeps the number of differences in each contig alignment under control by making corrections to differences that are likely due to sequencing errors. For each contig, DA reports a multiple sequence alignment, a profile, a consensus sequence, and the positions of popular differences in the contig. Experimental results on the seven largest clusters from NCBI UniGene and one large cluster from TIGR Gene Index show that DA can handle data sets of deep coverage efficiently. A comparison of the DA consensus sequences with the finished human genome indicates that the consensus sequences are of acceptable quality. The DA program can also generate splice variants.

Methods

A flow chart of the deep assembly algorithm is shown in Figure 1. Initially, a large word size is selected for computation of overlaps. The reads are considered one at a time for
computation of overlaps and construction of contigs. Whenever a good overlap involving the current read is found, a new contig is made of the two reads in the overlap or one of the two reads is added to an exiting contig containing the other read. When there are too many contigs or every read is considered at the current word size, the contigs are merged based on overlaps between contig consensus sequences, where overlaps are computed using the current word size. The steps given above are repeated for smaller word sizes until no contigs can be merged. Next, for each contig, a multiple alignment of reads is constructed, a final consensus and a profile are produced, and popular differences are generated based on the profile. Finally, the multiple alignments of the contigs are reported in ace file format, along with additional information on profiles and popular differences. Below we describe each step in detail.

**Computation of Overlaps**

Existing assembly programs first compute all overlaps between reads, then rank the overlaps, and use the overlaps in construction of contigs. However, it is not efficient to use this approach on a deep assembly problem because the number of overlaps between reads is quadratic in the number of reads. An alternative approach is to intertwine overlap computation with contig construction. To ensure that overlaps computed early are of strong similarity, we require that each overlap contains an exact word match of a large size. First we describe a method for computing overlaps between reads. Then we discuss a slight modification to the method for computing overlaps between contig consensus sequences.

A list of different word sizes in a decreasing order is used, one at a time, in computation of overlaps. Let $w$ be the current word size. The reads are processed one at a time to compute
overlaps between the current read and other reads. Exact word matches of size \( w \) between the current read and other reads are found by using a lookup table for a much smaller word size \( u \) (Huang and Madan 1999) and a hashing technique as follows. The word size \( u \) is selected such that there is enough memory to hold an array of \( 4^u \) integers. For every word \( x \) of length \( w \) in the current read, the word \( x \) is converted into a base-4 integer of \( w \) digits, and the integer is mapped by the hashing technique into an integer, denoted \( c(x) \), between 0 and \( 4^u - 1 \). The lookup table is used to locate all words \( y \) of size \( w \) in other reads such that \( c(y) = c(x) \). The words \( x \) and \( y \) are compared to see if they form an exact word match of size \( w \).

Each exact match is extended into a longest exact match if the longest exact match is not produced in previous extensions. The longest word match between the current read and another read corresponds to a diagonal in the dynamic programming matrix of Smith and Waterman (1981). For each longest match, a band of diagonals is formed by extending from the diagonal corresponding to the match in both directions by \( bd \) bases, where \( bd \) is a user-controlled parameter. An overlapping alignment of the two reads with the maximum score in the band is computed by a modified Smith-Waterman method (Huang 1992). The base quality values of the reads are used in the alignment computation (Huang and Madan 1999). The alignment is an overlap between the two reads. The overlap is partitioned into three regions: a 5' different end region, a similar region, and a 3' different end region, where the 5' (3') end region is a 5' (3') end region with the minimum score. The overlap is evaluated by two measures. One measure ensures that the length of the similar region divided by the length of the shorter read is greater than a cutoff. The other measure ensures that the length of each of the two different end regions divided by the length of the similar region is less
than a cutoff. If the overlap is good under both measures, then the overlap is immediately considered in construction of contigs.

The method described above is slightly modified for computing overlaps between contig consensus sequences. The modification is due to a difference between reads and contigs. Unlike reads, contigs are dynamically created and removed. As a consequence, the current set of contig consensus sequences is variable. Thus, the lookup table has to be kept up to date for the current set of contig consensus sequences. When two contigs are merged into a new contig, resulting in a new set of contig consensus sequences, a new lookup table is constructed for the new set of sequences.

**Construction of Contigs**

A contig is a group of reads that are oriented and aligned. A contig can be represented by a multiple alignment of reads. However, this representation is space-expensive because old contigs are removed and new contigs are created. A more space-efficient representation is used. Each contig is represented by a consensus sequence and two profiles: a base count profile and a quality value profile. For each position, the base count profile keeps the numbers of read bases of every type, the number of gap symbols, the number of 5' terminal bases, and the number of 3' terminal bases. The first and last bases of a read are 5' terminal and 3' terminal, respectively. Note that only non-terminal gaps, gaps flanked by bases in both directions, are counted. The numbers of 5' and 3' terminal bases in a profile are introduced for computing the numbers of gap symbols in another profile. For each position, the quality value profile keeps the sums of base quality values of every base type, the sum of quality values of gap symbols, the sum of quality values of 5' terminal bases, and the sum of quality
values of 3' terminal bases. The consensus sequence is generated from the two profiles. To save space, profiles are kept only for contigs of two or more reads.

When a good overlap between two reads is computed, the overlap is immediately considered in construction of contigs as follows. If none of the two reads in the overlap is in any contig, then a new contig is made from the two reads as follows. A list of the two reads is constructed for the new contig. If the 5' different end region of the overlap is not empty, then the shorter of the two read ends in the region is trimmed. The trimming procedure is also performed for the 3' different end region of the overlap. The resulting overlap consists only of a similar region and two terminal regions with each involving one read. A base count profile and a quality value profile are constructed based on the resulting overlap. A consensus sequence is generated from the profiles.

If one read is not in any contig and the other read is in a contig, then the read is aligned with the consensus sequence of the contig. If the alignment meets the requirements for good overlaps, then the read is added to the contig as follows. The alignment is cleaned by trimming different end regions of the read or the consensus sequence. If any end region of the consensus sequence is trimmed, then the end regions of the two corresponding profiles are also trimmed. For some positions of the resulting profiles, read bases that are internal before the trimming become 5' or 3' terminal bases after the trimming. Thus, the profiles are updated at the affected positions on the counts and quality values of terminal bases. The updating step is described in detail below. Base count and quality value profiles for the new contig are constructed from the updated profiles based on the clean alignment. The read list of the new contig is obtained by adding the read to the read list of the old contig.
If each of the two reads is in a different contig, then no action is immediately taken for the overlap. This case is to be addressed by the following section on merging contigs. Because it takes a substantial amount of space to represent a contig, it is necessary to keep an upper limit on the number of contigs. When the number of contigs reaches the limit or all reads are considered at the current word size in the computation of overlaps, contigs are merged as follows. Overlaps between contig consensus sequences are computed at the current word size. When a good overlap between consensus sequences is obtained, the overlap is immediately considered. The overlap is cleaned by trimming different end regions of one or the other consensus sequence. The profiles of the two contigs are properly trimmed and updated. Base count and quality value profiles for the new contig are constructed from the updated profiles of the two contigs based on the clean overlap. Figure 2 shows generation of a column of a new profile from two columns of old profiles for the case involving a gap. The read list of the new contig is obtained by merging the read lists of the two contigs.

We address a special problem with construction of deep contigs. As more reads are added into contigs, insertion sequencing errors result in gap positions in contig profiles. A position of a profile is a gap position if both the number of gap symbols and the proportion of gap symbols at the position are greater than cutoffs. Gap positions affect the accuracy of the method for constructing contigs. If the depths of coverage of a contig at every position are very high, gap positions occur between every pair of match positions in the profiles and consensus sequence. It is difficult to align accurately consensus sequences with 50% gap positions. An additional step is performed to deal with the gap problem. After a read is added into a contig or two contigs are merged, the gap positions in the new profile are removed. The resulting profiles are updated on the counts and quality values of terminal bases.
We describe how to update profiles after end regions of the profiles are trimmed or the gap positions of the profiles are removed. If the removed positions involve no terminal bases, that is, the counts of 5' and 3' terminal bases at the removed positions are equal to 0, no change is made to the profiles. Otherwise, a consensus sequence is generated from the profiles resulting from end trimming or gap removal. A subset of reads with terminal bases falling in the removed positions is located from the read list of the contig. Each read in the subset is aligned with the consensus sequence. If the read has no overlap with the consensus sequence, then the read is removed from the read list of the contig. Otherwise, a region \( R \) of the read that is aligned with the consensus sequence is located. The first and last positions of region \( R \) in the read, called the clipping positions of the read, are recorded for the read. If the read has a 5' terminal base in a removed position, then the position of the profiles aligned to the 5' terminal base \( b \) of region \( R \) is updated by increasing its 5' terminal base count by 1 and adding the quality value of base \( b \) to its 5' quality value sum. If the read has a 3' terminal base in a removed position, then the corresponding position of the profiles is updated.

Refinement of consensus sequences and profiles

In the contig construction step, each contig is represented by two profiles and a consensus sequence, instead of a multiple alignment, for space efficiency. After the construction step, the representation for each contig is refined as follows. For each contig, a multiple alignment is constructed by aligning each read in the contig with the consensus sequence of the contig. Then two new profiles and a new consensus sequence are constructed from the multiple alignment. The new consensus sequence is trimmed at ends if
the depth of coverage for the end is lower than a cutoff. The trimming based on the coverage information is useful for removing contaminants and chimeric ends of low coverage.

Popular differences in each contig are found based on the base count profile of the contig, which shows the percent occurrence of each base type at every position of the contig. A popular difference occurs at a position of the contig if the depth of coverage at the position is greater than a cutoff and there are at least two base types at the position, each with its percent occurrence greater than a cutoff. Popular differences are due to SNPs, sequencing errors in generation of ESTs, or mergers of ESTs from paralogous regions.

Results

The algorithm for assembling sequences of deep coverage was implemented as a computer program named DA (Deep Assembly). The DA program reports, for each contig, the layout of reads in the contig, a multiple alignment, two profiles, a consensus sequence, and the position of each popular difference in the contig. The multiple alignment is reported both in the ace file format for editing and viewing in Consed (Gordon et al. 1998) and in plain text format.

The efficiency and accuracy of DA were evaluated on eight data sets, each with over 9,000 EST sequences. All of the data sets were downloaded from UniGene database (Wheeler et al. 2003). The evaluation was performed on a Dell computer with 2 processors of 3.2 GHz and 4 GB of memory. Values for the parameters of DA were selected as follows. The maximum number of contigs was set to 400. Because this parameter controls the amount of space used for the contig data structures, the selection of the value 400 kept the amount of memory under 1 GB. The minimum percent identity of acceptable overlaps was set to 80%,
which was selected based on our experiences with CAP3 (Huang and Madan 1999). The minimum ratio of the overlap length to the length of the shorter read in the overlap was set to 75%. Because of deep coverage at each position, most of the true overlaps have a ratio greater than the cutoff. Many false overlaps have a ratio smaller than the cutoff. On each data set, the program was run in 3 iterations with word sizes 105, 75, and 45. The use of those large word sizes ensured that overlaps of high similarity were computed first.

**Efficiency of DA**

The overlap computation is the most time-consuming part of DA. The number of potential overlaps between sequences in a data set of deep coverage is very large. Thus the time efficiency of DA on a data set is measured by the ratio of the number of computed overlaps (alignments) to the number of sequences in the data set. The ratio ranged from 1.03 to 9.23 with an average of 4.98 for the eight data sets (Table 1). Compared with the maximum depths of coverage for the eight data sets shown in Table 1, the low ratios indicate that DA efficiently handled the eight data sets of deep coverage. The differences between the ratios are due to the types of overlaps in the data sets. If the overlaps are highly and entirely similar, the ratio is close to 1. On the other hand, if the overlaps are lowly or partially similar due to paralogous genes, alternatively spliced genes, or contaminations, the ratio is larger. This is because many overlaps are computed and rejected in the construction of contigs. As to be shown below, some of the eight data sets contained paralogous genes or alternatively spliced genes. Table 1 shows that the running time of DA is linear in the number of overlaps computed. It took about 0.13 second to compute an overlap. The space requirement of DA is about 30 times the size of the data set (data not shown). The DA program took at most 1 GB
of memory on each of the eight data sets. To test the ability of the program to handle a data set of deeper coverage, a synthetic data set with 100,000 sequences was generated. The DA program was run on the data set under the same set of parameter values. The program produced contigs with the maximum depth of coverage greater than 28,000 in 417 minutes and under 1.5 GB of memory.

**Accuracy of DA**

The accuracy of DA was evaluated by comparing DA assembly results with genomic sequences. Because complete high-quality genomic sequences are available only for the five human UniGene data sets (with names starting with 'Hs.'), the comparison is limited to the five human UniGene datasets. For each human data set, major contigs of coverage depth at least 100 produced by DA on the data set were selected for evaluation. For each major contig, the number of reads, the length of the consensus sequence, and the maximum coverage are shown in Table 2. The deepest contig in Table 2 has a coverage depth greater than 9,000. The number of popular differences for each major contig is also shown in Table 2, where the coverage depth cutoff was set to 100 and the percent occurrence cutoff was set to 10%.

The consensus sequence of each major contig was compared with the human genomic sequences with BLAT (Kent 2002). For each data set, the major contig consensus sequences from the set are similar to the same set of one to three genomic regions. For example, the five major consensus sequences from the data set Hs.439552 are all similar to the same set of three genomic regions: a region of Chr. 6, a region of Chr. 7, and a region of Chr. 9. The Chr. 6 region contains introns, whereas the Chr. 7 and Chr. 9 regions are free of introns. For each
consensus sequence, a best match between the consensus sequence and a genomic region was identified and the percent identity of the match is reported in Table 2. The percent identity of the best match for each consensus sequence ranges from 98.6% to 100%. However, some consensus sequences contain contaminants at ends, where a contaminant is not similar to any part of the human genome. The depth of coverage for an end of a contig with a contaminant is as high as 17. Because of the high depth coverage for contaminants, we have not found an effective method for finding and removing contaminants in the DA program. The lengths of contaminants at each end of every major contig are reported in Table 2. On three of the five data sets, DA could not merge some major contigs into one contig because the contigs contain different long contaminants at ends.

For each of the five data sets, there exist alternative splice forms of mRNA sequences. To see if DA could generate some splice variants, all mRNA sequences in the data sets were removed and each resulting data set with only EST sequences was given to DA as input. On the data sets Hs.14376 and Hs.426930, the major consensus sequences produced by DA cover two known alternative splice forms of mRNA sequences. On the other data sets, the major consensus sequences cover only one known splice form of mRNA sequences. The reason for missing other known alternative splice forms is that those forms are not represented by EST sequences of deep coverage. Note that DA keeps only contigs of sufficiently deep coverage. In addition, the contigs produced by DA contain several possible alternative splice variants that are different from the known mRNA sequences. All the variants were confirmed by comparing the contig consensus sequences with the human genomic sequences.
Discussion

We have developed the DA program to assemble EST sequences of very deep coverage. Experimental results produced by DA on large data sets of EST sequences show that DA is able to produce contigs of coverage depth in thousands in reasonable time. The technique used in DA for dealing with sequencing errors in EST sequences works very well to keep the level of difference in the contig alignment under control.

Popular SNPs in EST sequences can be found by using DA and by performing a post-assembly check. Each base type in a popular SNP is sufficiently represented by the population. The DA program produces a profile for each contig and reports the positions of popular differences in the contig. A post-assembly check is needed to select popular SNPs from the popular differences. Note that DA is not designed to find all SNPs. A method based on base quality values is very effective for finding SNPs in EST sequences of low coverage depth.

A major weakness in DA is its inability to find and remove contaminants of deep coverage at ends of contigs. A way to work around this weakness is to find and remove contaminants at ends of EST sequences through comparison with a database of contaminants. If a complete genome sequence is available, EST sequences can also be screened for contaminants by finding sequence ends that are not similar to any part of the genome sequence.

Acknowledgements
J. Wang and X. Huang were supported in part by NIH Grants R01 HG01502-05 and R01 HG01676-05 from NHGRI.

References


### Tables

#### Table 1: Performance of DA on eight data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>No. of reads (R)</th>
<th>No. of alignments (A)</th>
<th>No. of singletons</th>
<th>Running time (second) (T)</th>
<th>A / R</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC619155seqs</td>
<td>9193</td>
<td>9442</td>
<td>36</td>
<td>607</td>
<td>1.03</td>
</tr>
<tr>
<td>Cin.5491</td>
<td>10291</td>
<td>11144</td>
<td>37</td>
<td>838</td>
<td>1.08</td>
</tr>
<tr>
<td>Hs.14376</td>
<td>14652</td>
<td>92622</td>
<td>602</td>
<td>5777</td>
<td>6.32</td>
</tr>
<tr>
<td>Hs.169476</td>
<td>16135</td>
<td>82416</td>
<td>547</td>
<td>4802</td>
<td>5.11</td>
</tr>
<tr>
<td>Hs.406283</td>
<td>15336</td>
<td>74364</td>
<td>529</td>
<td>4433</td>
<td>4.85</td>
</tr>
<tr>
<td>Hs.426930</td>
<td>11850</td>
<td>109427</td>
<td>718</td>
<td>4831</td>
<td>9.23</td>
</tr>
<tr>
<td>Hs.439552</td>
<td>27847</td>
<td>229459</td>
<td>1164</td>
<td>11484</td>
<td>8.24</td>
</tr>
<tr>
<td>Mm.335315</td>
<td>11843</td>
<td>47290</td>
<td>209</td>
<td>2747</td>
<td>3.99</td>
</tr>
</tbody>
</table>

#### Table 2: Summary of accuracy information about major contigs produced by DA on five human data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>NAS(^a) Contig</th>
<th>No. of reads</th>
<th>Length</th>
<th>MDC(^b)</th>
<th>NPD(^c)</th>
<th>PI(^d)</th>
<th>LLPE(^e)</th>
<th>LRPE(^f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hs.14376</td>
<td>2</td>
<td>1</td>
<td>11480</td>
<td>2727</td>
<td>7309</td>
<td>66</td>
<td>99.9</td>
<td>445</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>142</td>
<td>1189</td>
<td>139</td>
<td>37</td>
<td>99.3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>1720</td>
<td>1142</td>
<td>1301</td>
<td>54</td>
<td>99.9</td>
<td>227</td>
</tr>
<tr>
<td>Hs.169476</td>
<td>1</td>
<td>1</td>
<td>14901</td>
<td>1985</td>
<td>9384</td>
<td>117</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>Hs.406283</td>
<td>1</td>
<td>1</td>
<td>14117</td>
<td>1977</td>
<td>8514</td>
<td>34</td>
<td>100</td>
<td>228</td>
</tr>
<tr>
<td>Hs.426930</td>
<td>2</td>
<td>1</td>
<td>6488</td>
<td>2228</td>
<td>2898</td>
<td>22</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>3412</td>
<td>923</td>
<td>3299</td>
<td>12</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>131</td>
<td>1120</td>
<td>125</td>
<td>33</td>
<td>100</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>116</td>
<td>1298</td>
<td>102</td>
<td>40</td>
<td>98.6</td>
<td>0</td>
</tr>
<tr>
<td>Hs.439552</td>
<td>1</td>
<td>1</td>
<td>20166</td>
<td>2461</td>
<td>8741</td>
<td>33</td>
<td>100</td>
<td>980</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>663</td>
<td>972</td>
<td>655</td>
<td>25</td>
<td>99.4</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>4136</td>
<td>1186</td>
<td>4109</td>
<td>90</td>
<td>98.7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22</td>
<td>110</td>
<td>743</td>
<td>110</td>
<td>18</td>
<td>99.1</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>116</td>
<td>809</td>
<td>116</td>
<td>40</td>
<td>98.7</td>
<td>48</td>
</tr>
</tbody>
</table>

\(^a\)NAS: Number of alternative splice forms. To test the ability of DA for generating alternative splice forms, it was run on the data set with only EST sequences. Data on the other columns are based on results by DA on the complete data with both EST and mRNA sequences.

\(^b\)MDC: Maximum depth of coverage

\(^c\)NPD: Number of popular differences (>10%)

\(^d\)PI: Percent identity

\(^e\)LLPE: Length of left poor end

\(^f\)LRPE: Length of right poor end
Figure Legends

**Figure 1:** A flow chart of the assembly algorithm.

**Figure 2:** Computation of a quality value profile column for the case involving a deletion gap. Column j of contig 2 is aligned to a gap immediately after column i of contig 1, based on an overlap between the consensus sequences of contigs 1 and 2. For quality value profile column i of contig 1, A_ii is the sum of quality values for base type A on column i of contig 1, C_ii is for base type C, G_ii is for base type G, T_ii is for base type T, D_ii is the sum of quality values for a deletion gap that occurs immediately after column i of contig 1, S_ii is the sum of quality values for 5' terminal bases on column i of contig 1, and E_ii is the sum of quality values for 3' terminal bases on column i of contig 1. The notations for column j of contig 2 are similarly defined. The deletion gap column is referred to as column k of the resulting contig. Quality value profile column k of the resulting contig is computed from quality value profile column i of contig 1 and quality value profile column j of contig 2.
Figures

Figure 1:

- Compute overlaps for a word size $w$
  - Otherwise
    - If all reads are considered
      - Merge contigs using the same method and word size $w$
  - If a good overlap is found, build a new contig or add one read into a contig
    - Refine consensus sequences and profiles
      - Find popular differences based on contig profiles
      - Report alignments, profiles and popular differences
  - Decrease $w$
- If $w$ is large

Otherwise


Figure 2:

\[
X = A_{ii} + C_{ii} + G_{ii} + T_{ii} + D_{ii} - E_{ii} + D_{ij}
\]
CHAPTER 3 A METHOD FOR FINDING SINGLE-NUCLEOTIDE POLYMORPHISMS WITH ALLELE FREQUENCIES IN SEQUENCES OF DEEP COVERAGE

Jianmin Wang and Xiaoqiu Huang

Abstract

Background

The allele frequencies of single-nucleotide polymorphisms (SNPs) are needed to select an optimal subset of common SNPs for use in association studies. Sequence-based methods for finding SNPs with allele frequencies may need to handle sequences with a coverage depth in thousands.

Results

We describe a computational method for finding common SNPs with allele frequencies in single-pass sequences of deep coverage. The method enhances a widely used program named PolyBayes in several aspects. We present results by our method and by PolyBayes on eighteen data sets of human EST sequences with deep coverage. The results indicate that our method used almost all single-pass sequences in computation of the allele frequencies of SNPs.

Conclusions

The new method is able to handle single-pass sequences of deep coverage efficiently. The method suggests that it is possible to analyze sequences of deep coverage without building a multiple sequence alignment of deep coverage.
Background

Information concerning the allele frequencies of single-nucleotide polymorphisms (SNPs) is needed to select an optimal subset of common SNPs for use in association studies [1]. One approach to finding common SNPs with allele frequencies is to generate DNA sequences from a sufficient number of samples in a population. This approach requires that computational methods have an ability to handle sequences of deep coverage (with a coverage depth in thousands). In this paper, we describe a computational method for finding common SNPs with allele frequencies in sequences of deep coverage. We present results produced by the method on human EST sequences of deep coverage, which are currently a major source of DNA sequences of deep coverage. The method is also expected to be useful for finding common mutations in sequences of deep coverage produced in a cancer genome project [2].

The PolyBayes program is widely used to find SNPs in redundant DNA sequences [3,4]. It first constructs a multiple sequence alignment based on pairwise alignments of each sequence with a high-quality genomic sequence called an anchor. Then it identifies and removes paralogous sequences that have a high number of observed differences with the anchor sequence. Next it computes an SNP probability score for each column of the multiple sequence alignment based on a rigorous Bayesian formula. The formula uses the prior probabilities of all the nucleotide permutations for the column, which are estimated from the quality scores of the bases on the column.

We enhance the PolyBayes program in several aspects to handle single-pass sequences (query sequences) of deep coverage. First, all the paralogous regions of the finished human
genome sequence are included as anchor sequences. Each query sequence is assigned to the corresponding anchor sequence that is different from each of the remaining anchor sequences at some positions but is identical to the query sequence at most of the positions. This approach separates paralogous sequences by making use of the positions where paralogous sequences differ but sequences from the same genome location agree.

Second, pairwise alignments of corresponding query and anchor sequences are used to construct profiles, one per anchor sequence. At each position of an anchor sequence, its profile contains the numbers and types of high-quality query bases that are aligned to the position of the anchor sequence. Candidate SNPs are produced based on the profiles, instead of multiple sequence alignments for the following reason. As the number of single-pass sequences in a multiple sequence alignment increases, the number of gap columns in the alignment increases but the number of identity columns in the alignment does not increase. Thus, it is difficult to construct an accurate multiple sequence alignment for single-pass sequences of deep coverage.

Third, because the pairwise alignment of corresponding query and anchor sequences may contain regions of low similarity due to sequencing errors or contaminants, the highly similar regions of the alignment are found by a dynamic programming algorithm. Only the highly similar regions are used in generation of the profile.

Our computer program named PolyFreq was compared with PolyBayes on eighteen data sets of human EST sequences of deep coverage. Results produced by PolyFreq and PolyBayes indicate that PolyFreq ran to completion and used almost all input sequences in computation of the allele frequencies of SNPs for every data set.
Results

The method for finding SNPs with allele frequencies was implemented as a computer program. The source code of the program is freely available for academic use [5]. The program takes as input a set of high-quality anchor sequences and a set of query sequences with quality scores. The set of anchor sequences includes all the paralogous regions of the genome for the set of query sequences. The anchor and query sequences are from the same species.

The program reports candidate SNPs in the anchor sequences. For each candidate SNP, the program reports its position in the anchor sequence, its local context in the anchor sequence, and base types with a frequency greater than a cutoff. The frequency of a base type is also given in a rational form with the number of query bases of the type as the enumerator and the total number of query bases as the denominator.

To evaluate PolyFreq, eighteen data sets of human EST sequences of deep coverage were constructed as follows. Eighteen clusters of human EST sequences, each containing at least 1,000 EST sequences with trace information, were randomly selected from the April, 2005 release of the UniGene database [6]. The eighteen UniGene clusters also contain EST sequences without trace information. For each of the eighteen UniGene clusters, an EST data set was obtained by selecting all EST sequences with trace information from the cluster. The set of quality score sequences for each of the eighteen data sets was produced from the trace information with Phred [7, 8]. Each of the eighteen data sets of full-length EST sequences without any masking was used as a query set.
For each query set, its set of anchor sequences was obtained by comparing the query sequences with the finished genome sequence at the BLAT web server [9]. By using stringent settings for BLAT, a set of two human anchor sequences was produced for each of three query data sets, and a set of one human anchor sequence was produced for each of the remaining query data sets. Each set of anchor sequences was screened for repeats with RepeatMasker [10].

The PolyFreq program was run on each pair of query and anchor sets. The PolyFreq program ran successfully to completion for each of the eighteen data sets. The following values were used for the parameters of the program: 50, minimum depth of coverage for each candidate SNP; 0.1%, minimum minor allele frequency; 5 bp, minimum perfect block length; 20, minimum base quality score in the perfect block; 90%, minimum percent identity for query-anchor alignments; 97%, minimum percent identity for the highly similar regions of query-anchor alignments.

Although PolyBayes was not designed to deal with data sets of deep coverage, we tested PolyBayes on the eighteen data sets of deep coverage to see how PolyBayes would behave on the data sets. Because PolyBayes takes only one anchor sequence, the corresponding anchor sequence was selected and given to PolyBayes for each data set.

On eight of the eighteen data sets, PolyBayes ran successfully to completion. On the remaining data sets, PolyBayes terminated abnormally without producing any output file after running for a few hours. The default values for all the parameters but the SNP probability output cutoff of PolyBayes were used; PolyBayes terminated abnormally more
frequently under other parameter values. A value of 0.75 for the SNP probability output
cutoff was used to produce a lower number of false positives than the default value of 0.5.

The abnormal termination of PolyBayes might be related to the deep coverage of the
data set and full-length EST sequences with low-quality ends or contaminants. Thus, for
each set of full-length EST sequences, a set of trimmed EST sequences was produced by
removing the ends of every sequence that are not highly similar to the corresponding anchor
sequence. For each data set, the number of trimmed sequences was almost equal to the
number of full-length sequences. The PolyBayes program was also run on each set of
trimmed sequences. It ran to completion for thirteen out of the eighteen data sets.

All the tests were performed on a Dell workstation with two 3.0-Ghz processors and 4
Gb of main memory. PolyFreq took less than one hour on every data set, whereas PolyBayes
was two to ten times slower than PolyFreq on every data set. The memory requirements of
PolyFreq and PolyBayes on the data sets were similar and were about 30 to 40 times the
input size.

The PolyFreq and PolyBayes programs were compared on every data set for which
PolyBayes ran successfully to completion on either the set of full-length sequences or the set
of trimmed sequences. For each data set, results produced by PolyFreq on the set of full-
length sequences were included in the comparison, whereas results produced by PolyBayes
on both sets of full-length and trimmed sequences were included. The SNPs from the dbSNP
database [11] that were mapped by the following method to the anchor sequences were used
as true SNPs for the comparison. Each SNP from dbSNP is specified by a local sequence
context. For each data set of EST sequences with a RefSeq sequence [12], each SNP from
dbSNP that occurs in the RefSeq sequence was determined by finding the exact occurrence of its sequence context in the RefSeq sequence. Each SNP from dbSNP in the RefSeq sequence was mapped to a corresponding anchor sequence by using a spliced alignment of the RefSeq and anchor sequences. Because three data sets had no RefSeq sequence, no SNPs from dbSNP were mapped to the anchor sequences for the data sets.

For each program on every data set with a RefSeq sequence, the number of true positives, the number of false positives, and the number of false negatives were computed. The number of true negatives was not collected because of its large value. Also reported were the number of sequences in the data set and the number of sequences that were used by the program to compute candidate SNPs. The comparison results are shown in Table 1.

The results in Table 1 indicate that PolyFreq could handle the data sets of full-length reads with problem regions and with very deep coverage. PolyFreq used 1,662 to 6,873 sequences on the six data sets for which PolyBayes terminated abnormally. On the data sets for which PolyBayes ran to completion, PolyFreq was similar to PolyBayes in the number of true positives and the number of false negatives, and is better than PolyBayes in the number of false positives. PolyBayes used significantly fewer sequences than PolyFreq on some of the data sets. Note that the ability to use as many sequences as possible is necessary for accurate computation of the allele frequencies of SNPs.

**Discussion**

We originally developed a method for assembling sequences of deep coverage. The method constructs multiple sequence alignments of deep coverage for contigs. The method has to deal with a large number of gap columns in the multiple sequence alignment. We later
agreed with one of the reviewers that it is not necessary to construct multiple sequence alignments for analysis of sequences of deep coverage. The reviewer also suggested that we focus on SNP analysis in sequences of deep coverage. Those suggestions motivated us to develop the method reported in this paper.

The PolyFreq program keeps PolyBayes' feature of performing comparisons between query and anchor sequences, instead of performing comparisons among query sequences. In addition, PolyFreq constructs profiles by using the highly similar regions of pairwise alignments of corresponding query and anchor sequences, instead of multiple alignments of query and anchor sequences. Thus, the efficiency and accuracy of PolyFreq are not significantly affected by query sequences of deep coverage. On the contrary, PolyFreq can compute the allele frequencies of SNPs more accurately in query sequences of deep coverage.

As sequencing costs are significantly reduced in the future, single-pass sequences from hundreds to thousands of individuals will be produced. Those sequences will be of deep coverage. Our current work suggests that it is possible to analyze sequences of deep coverage by using pairwise alignments of the sequences with the finished genome sequence, instead of multiple sequence alignments of deep coverage.

**Methods**

We first present the major steps of our method for finding common SNPs with allele frequencies in a set of query sequences and a set of anchor sequences. Then we describe each step in detail. The method consists of the following major steps:

1. Compute an alignment of anchor sequences for each pair of anchor sequences.
2. Compute an alignment of query and anchor sequences for each pair of similar query and anchor sequences.

3. For each query sequence, find the corresponding anchor sequence that is different from each of the remaining anchor sequences at some positions but is identical to the query sequence at most of the positions.

4. Find the highly similar regions of their alignment for each pair of corresponding query and anchor sequences.

5. For each anchor sequence, use the highly similar regions of every alignment involving the anchor sequence to construct a profile for the anchor sequence. At each position of the anchor sequence, its profile contains the numbers and types of high-quality query bases that are aligned to the position of the anchor sequence.

6. Report each candidate SNP with major and minor allele frequencies if its minor allele frequency is greater than a cutoff.

In step 1, for each pair of anchor sequences, an alignment of the sequences in given orientation as well as an alignment of the sequences in opposite orientation is constructed with GAP3, a global alignment program specially designed for genomic sequences with long different regions between similar regions [13]. One of the two alignments with a larger score is saved for the pair of sequences. The alignments saved in this step are to be used in step 3 for finding the corresponding anchor sequence for each query sequence.

In step 2, pairs of similar query and anchor sequences are found with DDS2, which produces a high-scoring chain of segment pairs (ungapped alignment fragments) between the two sequences in the pair [14]. For each pair of similar query and anchor sequences, an
alignment of the sequences in the pair is constructed with GAP22, an improved version of the GAP2 program [15] for quickly computing an alignment in a small area of the dynamic programming matrix, which is determined based on the chain of segment pairs. If the percent identity of the alignment is greater than a cutoff, then the alignment is saved for the pair of sequences.

In step 3, for each query sequence that is highly similar to two or more anchor sequences, the corresponding anchor sequence for the query sequence is selected among the anchor sequences through pairwise comparisons. Initially, one anchor sequence is taken as the current leader. Then the rest are compared with the current leader one at a time. Consider the comparison between the current leader and the current challenger. The winner between the two anchor sequences is produced by using the alignment of the two anchor sequences and their alignments with the query sequence. A common match occurs at a position of the query sequence, a position of the current leader, and a position of the current challenger if the three positions are pairwise aligned on each of the three alignments and contain the same base. The winner between the two anchor sequences is the one with a larger number of uncommon matches in its alignment with the query sequence. The winner becomes the current leader. After all the pairwise comparisons, the final leader is the corresponding anchor sequence for the query sequence.

In step 4, for each pair of corresponding query and anchor sequences, the highly similar regions of the alignment of the two sequences are identified in linear time with LCP, a program for finding regions of a sequence that meet a content requirement [16]. Each of the highly similar regions found by LCP has a percent identity greater than or equal to a cutoff
and is strictly optimal. The score of a region of the alignment is the sum of scores of every base match and every base difference in the region, where the score of every base match is \(1 - p\) and the score of every base difference is \(-p\). A region is optimal if its score is not less than the score of any other region that overlaps with it. An optimal region is strictly optimal if it is not completely contained in any optimal region other than itself.

In step 5, only substitutions in the highly similar regions of every alignment of corresponding query and anchor sequences are used to construct a profile for the anchor sequence because the remaining regions of the alignment have a high rate of difference, which is likely due to sequencing errors or contaminants in the query sequence. Additional requirements are introduced below because a long highly similar region may still contain a packet of sequencing errors in the middle. A sufficiently long section in a highly similar region of an alignment is a perfect block if the section consists only of exact base matches and the quality score of each query base in the section is greater than or equal to a cutoff [17]. A substitution in a highly similar region of an alignment is acceptable if it is immediately flanked on each side by a perfect block. An acceptable substitution is illustrated in Figure 1.

For each anchor sequence, its profile contains four counts at each position: one count for each query base type. For example, the base A count at the anchor position is the number of acceptable substitutions at the anchor position and at any query sequence position with base A, in the highly similar regions of the alignments of the anchor sequence with corresponding query sequences. A count for a query base type at the anchor position is 0 if there is no acceptable substitution at the anchor position and at any query sequence position with the
query base type. The frequency of each of the four counts is the count divided by the sum of the four counts if the sum is positive.

In step 6, each profile is scanned for candidate SNPs. A candidate SNP occurs at an anchor sequence position if the sum of the four counts for the position is greater than or equal to a cutoff and at least two of the four counts have a frequency greater than a cutoff. All candidate SNPs with allele frequencies are reported along with a local anchor sequence region for each candidate SNP.

Author’s contributions

XH designed the strategy for solving the problem and provided guidance to JW. JW worked out the details of the strategy, developed the program, and produced results on data sets with the program. XH wrote the paper and JW formatted it in Word. All authors read and approved the final manuscript.

Acknowledgments

We thank Geo Pertea and John Quackenbush for discussions on assembly of sequences of deep coverage, and Brian Haas for suggestions on and evaluation of GAP22. We are grateful to the reviewers for suggestions that motivated us to develop the new method.

JW and XH were supported in part by NIH Grants R01 HG01502-05 and R01 HG01676-05 from NHGRI.

References


5. PolyFreq Program [http://bioinformatics.iastate.edu/aat/PolyFreq/].


The mark N/A means that no SNP from dbSNP was mapped to the anchor sequence because of lack of a RefSeq sequence. The mark T/A means that PolyBayes terminated abnormally without producing any output file. A candidate SNP from the program is considered as true positive (TP) if it is in dbSNP or false positive (FP) otherwise. A SNP from dbSNP that occurs in the data set is considered as false negative (FN) if it is not reported as a candidate SNP from the program. The number of sequences used (NSU) by the program in generation of candidate SNPs is reported.
**Figure Legends**

**Figure 1:** An acceptable substitution in an alignment. The line shows an alignment of query and anchor sequences with thick parts indicating highly similar regions. The small box contains an acceptable substitution that is flanked by a perfect block on each side, and the large rectangular box gives a detailed view of the small box. The quality value of each base in the large box is shown next to the base.
Figures

Figure 1:

Acceptable substitution

\[
\begin{array}{cccccccccc}
30 & 35 & 40 & 35 & 30 & 40 & 35 & 30 & 45 & 30 & 40 \\
T & G & A & T & C & A & G & A & T & C & A \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
T & G & A & T & C & T & G & A & T & C & A \\
35 & 40 & 30 & 35 & 40 & 30 & 35 & 40 & 30 & 45 & 30 \\
\end{array}
\]
CHAPTER 4 EVOLUTION OF ALTERNATIVE SPLICING AFTER GENE DUPLICATIONS

Zhixi Su¹, Jianmin Wang², Jun Yu, Xiaoqiu Huang, and Xun Gu

¹: these authors have equal contributions

Abstract

Alternative splicing and gene duplication are two major sources for proteomic function diversity. Here, we study the evolutionary trend of alternative splicing after gene duplication by analyzing the alternative splicing differences between duplicate genes. We observed that duplicate genes have less alternative splice (AS) forms than that of single-copy genes, and a negative correlation between the mean number of AS forms and the gene family size. Interestingly, we found that the loss of alternative splicing in duplicate genes may occur shortly after the gene duplication. These results support the subfunctionization model of alternative splicing in the early stage after gene duplication. Further analysis of the alternative splicing distribution in human duplicate pairs showed the asymmetric evolution of alternative splicing after gene duplications, i.e., the AS forms between duplicates may differ dramatically. We therefore conclude that alternative splicing and gene duplication may not evolve independently. In the early stage after gene duplication, young duplicates may take over a certain amount of protein function diversity that previously was carried out by the alternative splicing mechanism. In the late stage, the gain and loss of alternative splicing seem to be independent between duplicates.


**Introduction**

Alternative splicing, which was discovered decades ago, is a common post-transcriptional process in eukaryotic organisms to produce multiple transcript isoforms from a single gene (Black 2003). Although substantial evidence has shown the functional importance of alternative splicing in development, differentiation and cancer (Lopez 1998; Venables 2002; Jiang et al. 2000), alternative splicing was conventionally thought of as an exceptional event occurring in only 5% of human genes (Sharp 1994). However, such view has been shown incorrect by the genomic data. Indeed, many studies have revealed a very different picture that more than 50% of genes in human or mouse genes are alternatively spliced (Mironov et al. 1999; Brett et al. 2000; Kan et al. 2001; Brett et al. 2002; Kim et al 2004). Though the estimation of AS forms has been rough, which may vary among different approaches and EST datasets, it has been generally accepted recently that alternative splicing may play as one major mechanism for generating proteomic complexity in higher eukaryotes (Graveley 2001; Maniatis and Tasic 2002; Kriventseva et al. 2003).

The finding of high percentage alternatively spliced genes in the human and mouse raises several interesting evolutionary questions. For instance, gene duplications have been widely proposed as the major resources for the origin of new genes to increases the proteomic complexity by the follow-up functional divergence (Ohno 1970; Hughes 1994; Li 1997). So, what would happen when an alternatively-spliced gene is duplicated? Apparently, each duplicate copy could lose some AS isoforms due to the functional redundancy, or acquire new isoforms. Yu et al (2003) found that two duplicates of Fugu synapsin-2 genes, SYN2a and SYN2b, corresponded to each of AS isoforms of the single human SYN2 gene. Using the model of subfunctionization (Force et al. 1999), the authors suggested that the
ancestral Fugu gene prior to duplication may have two AS isoforms, each of which was kept by the duplicate gene that lost the potential to produce the other AS isoform. Another example is the teleost mitf duplicate genes (mitfa and mitfb). Altschmied et al. (2002) found that the long indels in the 5'-terminal might be the result of the complementary degeneration of alternative 5' exons of the ancestral gene after gene duplication.

In spite of these interesting case-studies that indicate gene duplications may reduce the level of alternative splicing, it remains largely unclear at the genome level, how alternative splicing evolves after gene duplications. Since both mechanisms are important for the evolution of functional diversity, one may raise some intriguing questions. For instance, how fast it may evolve to generate alternative splicing difference between duplicates? Is alternative splicing evolution asymmetric? Or whether alternative splicing and gene duplication evolve independently. To address issues, we develop a computational pipeline to predict the number of AS isoforms for each human gene, which is used as the proxy of the alternative splicing functional divergence. We then conduct a comprehensive analysis to investigate the evolutionary pattern of alternative splicing after gene duplications.

**Results**

**Detecting human Alternative splice forms**

Our interest here is to identify all potential AS forms for well-annotated human genes. To this end, we developed a computational pipeline to predict the number of AS isoforms for each human gene by comparing human ESTs and mRNAs to the fine-assembled human genome sequence (see Methods for a brief description of the pipeline). We have examined 15,422 non-redundant human genes (RefSeq) and found that 12,014 of them (77.9%) are
possibly alternatively spliced, that is, they may have at least two forms of message RNAs, or at least one AS forms. The average number of AS forms is 3.9 per human gene. However, the range of AS forms is unexpectedly broad among genes. For instance, there are 1,167 genes (7.6 %) that have more than 10 AS forms. We found that the number of genes decays with the number of AS forms, following an exponential law perfectly (Fig.1, R=0.99). That is, the frequency \( f(k) \) of \( k \)-AS forms can be characterized by a geometric distribution, \( f(k) = p(1-p)^k \), where \( k=0, 1, 2,...; \) the parameter \( p \) was estimated as \( p = 0.26 \).

**Duplicates may have less Alternative splice forms than single-copy genes**

Gene duplications are widely believed to be the major source of genetic novelties (Ohno 1970). Meanwhile, numerical examples have shown that combinations of alternative splicing in specific genes can significantly expand the coding capacity of genome, such as cell adhesion molecules or ion channels (Lopez 1998; Copley 2004). The “independent model” claims that alternative splicing and gene duplication are two independent mechanisms for increasing the proteomic complexity. Alternatively, the “function-sharing model” claims that some proteomic components can be performed either by alternatively spliced genes or duplicate genes. We have noticed that these two competing models have distinct predictions about the level of AS forms. For instance, the “independent model” predicts a similar level of AS forms between single-copy and duplicate genes, while the “function-sharing model” predicts a high level of alternative splicing in single-copy genes than that in duplicate genes, because AS forms can be fixed in each copy, respectively, after gene duplication (Fig.2).

Yu et al. (2003) conducted a few cases-studies to support the “function-sharing model”. To further test whether it is the general case, we used the BLAST search to classify all well-
annotated human genes under study into 8819 single-copy genes and 6603 duplicate genes (see Methods). The percentage of alternative splicing of single-copy genes is \( \frac{7090}{8819} = 80.4\% \), while that of duplication genes is given by \( \frac{4924}{6603} = 74.6\% \). The difference between single-copy and duplicates is statistically highly significant (chi-square = 74.3, \( p<10^{-5} \)). Similarly, the mean number of alternative splicing duplicate genes is \( 3.52\pm0.05 \), which is significantly lower than that of single-copy genes \( 4.11\pm0.05 \) (\( t \)-test, \( P<0.001 \)). Besides, we found a higher proportion of single-copy genes with many AS forms (8.3% for more than 10 forms) than that (6.6%) of duplicates. In short, our analysis suggests that, at the genome level, functional divergence among duplicate genes may reduce the number of AS forms, which is more consistent with the prediction of the “function-sharing model” rather than the “independent model”. Roughly speaking, one duplicate gene may take over a certain amount of proteomic diversity that previously were carried out by, on average, 0.59 (\( = 4.11-3.52 \)) AS forms. In other word, in the human genome, there are about \( 6603\times0.59 = 3896 \) losses of AS forms due to gene duplications.

**Large gene families tend to have less Alternative splice forms**

Another prediction from the “function-sharing model” is that large gene families tend to have less AS forms, because multiple-round of gene duplications may results more loss of AS forms. Except 1,486 genes that can not mapped to Ensembl gene ID or did not assigned Ensembl gene family ID, we further classified remaining 13,936 human genes into 8,211 gene families (see Methods), and for each gene family size, estimated the mean of AS forms and the percentage of genes that are not alternatively spliced (percentage of no-AS). Fig.3 (A) shows the mean of AS forms plotting against the gene family size; a single-copy gene is regarded as a gene family with size 1. It appears that the mean of AS forms remains roughly
constant for gene families with small to moderate sizes, say, size one to 4, but decreases for larger gene families \( (R = -0.85, p < 0.0037) \). Similarly, Fig.3 (B) shows that the proportion of no AS forms increases with the increasing of gene family size \( (R = 0.93, p < 0.0004) \). In short, the negative correlation between the number of gene AS forms and the gene family size is likely to be caused by the effect of function-sharing after gene duplications.

**Loss of alternative splicing may occur only shortly after gene duplication**

To further explore the evolutionary pattern of alternative splicing after gene duplications, we compiled independent 2875 human duplicate pairs, and calculated the protein sequence distance \( (d) \) between duplicates (Methods). As shown by the histogram (Supplemental Figure 1), there is a peak of duplicate pairs around \( d=0.7 \), roughly corresponding to 500-600 million years ago when the duplication time was estimated by the method of molecular clock (Gu et al. 2002a).

Using the protein sequence distance as a proxy of the age of duplicate, we first grouped duplicate gene pairs with similar protein sequence distance (with a bin of 0.1 distance unit, about 80 million years ago, or the time of mammalian radiation), and calculated the percentage of no-AS. Interestingly, the percentage of no-AS in the most recent duplicate group \( (d<0.1) \) is 42%, which is almost 2-times higher than that (22% on average) in the other more ancient groups (Fig.4 panel A); a chi-square test showed the difference is statistically significant \( (P<0.001) \). Similarly, the mean number of AS forms (2.7) in the most recent duplicate genes is smaller than that (3.5) of more ancient duplicate genes, as shown in Fig 4 (B). Hence, both measures, the percentage of no-AS, and the mean number of AS forms, indicated that a rapid reduction of AS forms in young duplicates.

**Duplication v.s. alternative splicing in other model organisms**
The above analysis was based on human AS isoforms. It would be interesting to test whether duplicates in other organisms also tend to have a lower number of AS isoforms. We developed a simple approach to compare the evolutionary patterns of alternative splicing in several model animals, and found it might be the case (Fig. 5). First, we used the amino acid identity percentage ($I$) and the Blast search E-value as the criteria to define duplicate genes. For example, the identity $I=30$ and E-value $1E-10$ mean genes have identity greater or equal to 30% and E-value is less or equal to $1E-10$. Second, we changed the criterion $I$ from 30% to 50%, 70%, and 90%, respectively, and calculated the corresponding proportion of no-AS genes. Since the amino acid identity ($I$) is a proxy for young and ancient duplicate genes, the proportion of no-AS genes is expected to increase with the increasing of $I$, if the loss of AS isoforms in duplicates may occur shortly after the gene duplication. For the comparison, we also included the human genome. As shown in Fig. 5, we found that in all organisms we examined (human, mouse, *Drosophila* and *C.elegans*), the proportion of no-AS genes tends to increase with the sequence identity criterion. Our analysis indicated that the loss of AS forms in young duplicates may be a general pattern in animals.

**Alternative splicing of duplicates after the human-mouse split**

To further test the hypothesis that loss of alternative splicing may occur only shortly after gene duplication, we classified all human genes into two groups using mouse orthologs as the reference. The A-group (H1) includes 9640 one-to-one and 282 one-to-many human-mouse orthologous genes, while the B-group (Hx) includes 419 human duplicate genes that were duplicated after the human-mouse split. We analyzed the alternative splicing evolution for the B-group, i.e., human recent duplicates after the human mouse split. We found that the mean AS forms of Hx group (2.83±0.2) is significantly lower than that of H1 group.
(3.66±0.04), as well as that of whole human duplicate genes (3.52±0.05). Similarly, the percentage of no-AS forms in Hx (0.369±0.024) is significantly higher than H1 group (0.221±0.0042). These results indicate that the reduction of AS forms in duplicate genes mainly happened in the early stage after the gene duplication.

Putting together, we propose a scenario for the evolution of alternative splicing after the gene duplication (Fig.2). For simplicity, consider that a gene has two AS forms, $L$ (long) and $S$ (short), for distinct physiological roles. After the gene/genome duplication, two duplicate copies ($A$ and $B$) inherit both AS forms, but may start to have differential expression profiles. Because of the functional redundancy, the transcription of the $L$-form in gene $A$ becomes more dominant, while that of $S$-form in gene $B$ becomes dominant. This evolutionary transition from alternative-splicing diversity to functional divergence of duplicates may occur in the early stage after gene duplication. In the late stage, novel alternative splicing may added, increasing the overall number of AS forms.

**Testing asymmetric evolution of alternative splicing after gene duplication**

Although there are 571 (19.8%) and 575 (20%) duplicate pairs with no or only one difference, respectively (Fig.6), we indeed observed a significant portion of duplicate pairs showing dramatic differences in their number of AS forms. For instance, 206 (7.2%) duplicate pairs were found to have more than 10 differences of AS forms, indicating the possibility of asymmetric pattern of alternative splicing evolution after gene duplication. It implies that not all duplicate genes have similar number of AS forms. To address this issue vigorously, we used the binomial test to obtain the $p$-value (type-one error) for each duplicate pair under the null hypothesis of no difference in the number of AS forms. As a result, 418 (14.5%) pairs have significantly different AS forms ($p<0.05$), and 181(6.3%) pairs have $p$-
value less than 0.01 (Fig.7). We showed here 36 duplicate pairs have the largest AS form
difference in table 1 (p<0.0001) (for all 181 pairs see Supplementary Table online). However,
since it involves 2875 simultaneous statistical tests, the multiple-test problem should not be
neglected. For instance, at the significance level of 0.05, there are about 2875×0.05=144
significant cases by the pure chance. In other words, at this 0.05 significance level we
observed 418 significant cases, only 418-144=274 cases are likely to be truly significant;
others are false-positive. Statistically, it may be evaluated by the false-positive discovery rate
(FDR). At the 0.05 significance level, we calculated FDR=34%. For 181 significant cases at
the level of 0.01, the false-positive discovery rate (FDR) is about 16%. At any rate, this
analysis indicates that after a significant portion of human duplicate genes evolve
asymmetrically in the AS forms.

The effects of EST coverage and expression level

Clearly, AS form detecting is affected by the EST coverage, because the more ESTs
found for a given gene, the more likely AS form(s) can be detected (Hide et al. 2001; Kan et
al. 2001). When the size of library is sufficiently large, the EST coverage of genes is mainly
determined by the gene expression levels. Indeed, we observed a positive correlation between
the detected AS form number and the EST hit number in our dataset (p<0.0001,
Supplemental Figure 2A). In addition, the number of genes decays exponentially with the
number of AS forms (Fig.1), as well as the number of EST hits (R = 0.93, Supplemental
Figure 2B).

We have run several tests to examine whether our duplicate-AS analysis is affected by
the EST coverage, measured by the number of ESTs alignmmented to a given gene (EST hits)
(see Methods). We found a very similar EST-hit distribution between single-copy and
duplicate genes; the mean number of EST hits in single-copy genes is 196±3.7, with no significant difference from that in duplicate genes (206.9±6.0). Moreover, we grouped duplicate genes and single-copy genes with similar EST coverage (with a bin of 40 EST hits), and found duplicate genes always have larger percentage of no AS genes and smaller mean of AS forms in any EST coverage span (Supplemental Figure 3).

In the AS-gene family size and AS-duplicate distance analyses, we classified all human genes into two categories: the $H$-category for 7713 highly expressed genes with more than 112 EST hits, and the $L$-category for 7709 other lowly expressed genes (less than 112 EST hits). We found the positive correlation between the proportion of no AS genes and gene family size in both $H$ and $L$ categories (Supplemental Figure 4). For the $L$-category, the proportion of no AS is decreased with the increasing of the duplicate protein distance. However, in the $H$-category, the proportion of no AS in single-copy genes and duplicate genes is decreased with the increasing of the duplicate protein distance when $d<0.3$, but seems to increase in more ancient duplicate pairs (Supplemental Figure 4). Though this pattern needs to be elaborated further, it may imply that highly expressed duplicate genes may lose some AS forms in the late stage after gene duplication.

Finally, we compared the duplicate pairs that have significant asymmetric evolution of AS forms (see Table 2). For 181 duplicate pairs with highly significant asymmetric AS forms ($P<0.01$), there are 98 (54%) pairs have the similar EST coverage (both belong to $H$ or $L$-category). Though it is reasonably (but not much) lower than that for the rest of 2694 duplicate pairs with $P>0.01$ (67%), it indicates that the asymmetric evolution of alternative splicing after gene duplication may not merely because of the gene expression divergence difference.
Discussion

We conducted a large-scale alternative-splicing analysis of the human genome, using human AS forms predicted from the EST databases. Many methods or programs for identification of alternative splicing were developed recently, such as mapping ESTs onto mRNA sequences (Brett et al 2002), aligning ESTs with genomic sequences (Mironov et al 1999), multiple alignment of ESTs, mRNAs and genomic sequences (Modrek et al 2001, Modrek and Lee 2003) and aligning transcribed consensus sequences to genomic sequences (Gupta et al 2004). A number of studies have relied on EST self-clustering to assemble alternatively spliced transcripts (Mironov et al. 1999; Quackenbush et al. 2001). Some studies have combined genome-based EST clustering and transcripts assembly approach to reconstruct the alternatively spliced isoforms (Kan et al. 2001; Xing et al. 2004; Kim et al. 2005a). All these afford attempt to address the 'garbage EST' issue. The analysis pipeline we developed has adopted several techniques to reduce the potential garbage EST effects (also see Methods). First, a simple method was implemented to verify splice forms based on the pairwise alignments between EST/RefSeq and genomic sequences to alleviate the error-prone nature of EST consensus sequences caused by high sequencing error rates in ESTs, experimental contaminations, chimeric clones, redundant copies, paralogous genes or pseudogenes. Second, any EST that results in an extremely short exon or intron is removed because extremely short exons and introns are error-prone. Third, splice-site motifs (GT-AG/GC-AG) were used to validate our detected AS forms. And finally, to avoid the high false positive rate of transcript assembly procedure (Bouck et al. 1999; Xing et al. 2004), we did not assemble the ESTs to full-length transcripts. Instead, we simply detected the number
of mutually exclusive splice, using this as a proxy for the level of AS-related functional diversity.

We found that the percentage of alternatively spliced genes and average number of AS forms per gene we estimated in the human genome is somewhat higher than previous prediction, e.g., see Kim et al. (2004). The difference may be from the different sampling strategies adopted. To obtain a more precise evaluation of alternative splicing, we selected the well annotated human genes in RefSeq database (Pruitt et al. 2005). These well annotated genes obviously have more high ESTs coverage than other genes.

Many non-functional alternatively spliced transcripts may be produced during pre-mRNA splicing (for review, see Modrek and Lee 2002; Lareau et al. 2004), such as aberrant splicing in some tumors, genome contamination, unspliced mRNA, splicing errors arose from various reasons, or background splicing without any effect on the cellular function. Some of these artifacts can be detected by identifying the premature termination codon (PTC) in alternative splicing isoforms to find the putative targets of nonsense-mediated decay (NMD) (Lewis et al. 2003; Xing and Lee 2004). In this current study, we were not able to use such approach because we did not assemble the EST to full-length transcripts. Instead, we excluded these garbage ESTs by requiring pairs of mutually exclusive splices in different ESTs. Since observing a given splice form in one EST may be insufficient, we tested a stricter criterion by requiring that the two ESTs share one splice site but differ at RefSeq. As expected, we detected about half AS forms than before, but our analysis showed it did not affect our main conclusions (Supplemental Figure 5).

In spite of these precautious treatments, inevitably, the AS forms detecting still has a certain level of error rate. To test whether our main results are not sensitive to the inherited
bias in our analysis pipeline, we repeated our analysis by using an independent predicted alternative splicing data derived from ECgene database (Version 1.2) (http://genome.ewha.ac.kr/ECgene/) (Kim et al. 2005a; Kim et al. 2005b). As shown in Supplementary figures (Supplemental Figure 6), we obtained virtually similar results.

**Conclusive Remarks**

In this article, we discovered the level of alternative splicing in duplicate genes is usually lower than that of single-copy genes. Further analyses indicated that the number of AS forms is negatively associated with the gene family size, and the loss of AS forms may happen shortly after the gene duplication. Moreover, we demonstrated the asymmetric AS evolution, i.e., the AS numbers between some duplicates may differ dramatically. Some of results have been confirmed in other organisms, including mouse, *drosophila* and *C.elegans*. We therefore conclude that AS and gene duplication, two mechanisms for proteomic function diversity, may not evolve independently, supporting *the model of function-sharing*. That is, in the early stage after gene duplication, the young duplicates may take over a certain amount of protein function diversity that previously was carried out by the alternative splicing mechanism. After that, the evolution of AS forms may be independent between duplicates.

Several studies showed that evolutionary rate of coding sequences may accelerate shortly after gene duplication (Lynch and Force 2000; Kondrashov et al. 2002; Conant et al. 2003; Zhang et al. 2003). Gu et al (2005) found that the initial rate for either expression or regulatory network evolution after yeast gene duplications is much higher than that of late stage. The current study for the loss of AS forms between duplicates presents another example to support the viewpoint of rapid evolution in the early stage after gene duplication. Moreover, the pattern of rapid AS form loss in young duplicates is consistent to the model of
subfunctionalization (Lister et al. 2001; Altschmied et al. 2002; Yu et al. 2003). We suggested that AS subfunctionalization of duplicate gene may be a general phenomenon, which had happened in the early stage after gene duplication. On the other hand, alternative splicing may contribute to the neofunctionalization by acquiring new functional AS variants, resulting in the increase of AS isoforms for ancient duplicate genes (Fig. 2).

Methods

Sequence Data

The reference sequences (RefSeq) of human, mouse, drosophila and C.elegans were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/refseq/). Only the mRNA sequences of IDs starting with the prefix “NM_” were extracted and the corresponding protein sequences were also extracted. There were 21,267, 16,863, 18,648 and 20,785 mRNA reference sequences for human, mouse, drosophila and C.elegans respectively. The alignments between mRNA reference sequences and genome sequence were downloaded from the UCSC genome center (http://genome.ucsc.edu/) and they were generated by BLAT (Kent et al. 2002) program. The alignment information was used to remove redundant sequences. In case of multiple mRNA sequences being mapped to one genomic region, only the sequence resulting in the longest protein product was accepted.

Identification of Duplicate Genes

We used the method of Gu et al (2002b) to identify duplicate genes. Briefly, every protein was used as the query to search against all other proteins by using Blastp ($E = 10$). Two proteins are scored as forming a link if (1) the Blastp alignable region between them is >80% of the longer protein, and (2) the identity ($I$) between them $I \geq 30\%$ if the alignable region is longer than 150aa and $I \geq 0.01n + 4.8L^{0.32(1+\exp(-L/1000))}$ (Rost 1999) for all other
protein pairs, in which \( n = 6 \) and \( L \) is the alignable length between the two proteins. Duplicate pairs were seeded with a two-way best pairwise match. The protein sequence distance \( (d) \) between duplicates was calculated by \( d = -\ln(L/100) \), the Poisson correction.

**Identifying Gene Family Size and Genes Duplicated After Human-Mouse Split**

All human gene family IDs were extracted using the EnsMart tool (Kasprzyk et al. 2004). The size of each gene family was obtained from counting the frequency of this gene’s family ID in whole genome. Orthologs and in-paralogs between human and mouse were extracted from Inparanoid database (version 4.0) (O’Brien et al. 2005). Human genes have one-to-one and one-to-many relationships with mouse were classified into HI category. Similarly, human genes have many-to-one and many-to-many relationships with mouse were classified into Hx category. RefSeq/Ensembl gene mapping was extracted using EnsMart, only reciprocally unique pairs were further analyzed.

**Identification of Alternative Splice Forms**

Alternative splice forms are identified by using alignments of human EST/RefSeq and genomic sequences as follows. The EST/RefSeq sequences that are highly similar to only one genomic region are selected. The remaining EST/RefSeq sequences are discarded. The selected EST/RefSeq sequences are grouped together by genomic regions, one group per genomic region. If there are multiple RefSeq sequences in a group, then the RefSeq sequence with the longest protein sequence is the leader sequence for the group. For each EST/RefSeq sequence, splice sites in the genomic region are identified by the GeneSplicer program (Pertea et al. 2001), and exons and introns of sufficient lengths in the genomic region are identified based on the splice sites and the alignment of the EST/RefSeq sequence and the genomic region. Any EST/RefSeq sequence without any exon and intron is discarded. The
pattern of splicing for each EST/RefSeq sequence is indicated by a list of all exons in order of increasing coordinate. For each group of EST/RefSeq sequences, alternative splice forms are identified by comparing the exon list of each EST/RefSeq sequence with the exon list of the leader sequence and finding differences in exon coordinates. Redundant alternative splice forms are removed.

The method given above finds and removes garbage ESTs from aberrant transcripts or abnormal cell lines by three measures. First, any EST that is not highly similar to a genomic region over a majority of its length is removed. Second, any EST that results in an extremely short exon or intron is removed because extremely short exons and introns are error-prone. Third, any EST that results in a weak splice site in a genomic region is removed.

Additional methods for finding garbage ESTs are suggested by several previous studies (Modrek and Lee 2002; Xing and Lee 2004; Lareau et al. 2004). One of them is to use only AS forms that are confirmed by two independent libraries. We are currently experimenting with this method and other methods to see if they affect the main observations in the paper.

Acknowledgements

This work was supported by the NIH grants (X.G.).

References


**Web site references**


http://inparanoid.cgb.ki.se; INPARANOID program and database. (date accessed: Feb, 2005)

### Table 1: Duplicate Pairs That Have the Highest Alternative Splicing Divergence

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>AS forms of gene 1</th>
<th>Gene 2</th>
<th>AS forms of gene 2</th>
<th>Function of gene 1</th>
<th>Function of gene 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_014364</td>
<td>2</td>
<td>NM_002046</td>
<td>63</td>
<td>GAPDH-2</td>
<td>glyceraldehyde-3-phosphate dehydrogenase, GAPDH</td>
<td>0</td>
</tr>
<tr>
<td>NM_000900</td>
<td>57</td>
<td>NM_000393</td>
<td>2</td>
<td>collagen, type III, alpha 1</td>
<td>collagen, type V, alpha 2</td>
<td>1.5E-14</td>
</tr>
<tr>
<td>NM_093426</td>
<td>46</td>
<td>NM_002073</td>
<td>3</td>
<td>GNAS complex locus</td>
<td>G protein, alpha z polypeptide</td>
<td>1.74E-13</td>
</tr>
<tr>
<td>NM_001658</td>
<td>39</td>
<td>NM_001659</td>
<td>3</td>
<td>ADP-ribosylation factor 1</td>
<td>ADP-ribosylation factor 3</td>
<td>8.53E-09</td>
</tr>
<tr>
<td>NM_003970</td>
<td>38</td>
<td>NM_003803</td>
<td>3</td>
<td>myomesin (M-protein) 2</td>
<td>myomesin 1 (skelemin)</td>
<td>1.55E-08</td>
</tr>
<tr>
<td>NM_184041</td>
<td>43</td>
<td>NM_005165</td>
<td>6</td>
<td>aldolase A</td>
<td>aldolase C</td>
<td>6.06E-08</td>
</tr>
<tr>
<td>NM_003127</td>
<td>30</td>
<td>NM_003126</td>
<td>1</td>
<td>α-spectrin, non-erythrocytic 1</td>
<td>α-spectrin, erythrocytic 1</td>
<td>6.54E-08</td>
</tr>
<tr>
<td>NM_003971</td>
<td>3</td>
<td>NM_033392</td>
<td>35</td>
<td>sperm associated antigen 9</td>
<td>protein kinase binding</td>
<td>9.29E-08</td>
</tr>
<tr>
<td>NM_002473</td>
<td>30</td>
<td>NM_005964</td>
<td>2</td>
<td>myosin, heavy polypeptide 9</td>
<td>myosin, heavy polypeptide 10</td>
<td>3.83E-07</td>
</tr>
<tr>
<td>NM_000188</td>
<td>23</td>
<td>NM_000189</td>
<td>0</td>
<td>hexokinase 1</td>
<td>hexokinase 2</td>
<td>7.75E-07</td>
</tr>
<tr>
<td>NM_006087</td>
<td>1</td>
<td>NM_178014</td>
<td>25</td>
<td>tubulin, beta 4</td>
<td>tubulin, beta polypeptide</td>
<td>1.52E-06</td>
</tr>
<tr>
<td>NM_012268</td>
<td>25</td>
<td>NM_138790</td>
<td>1</td>
<td>phospholipase D3</td>
<td>hydrolase activity</td>
<td>1.52E-06</td>
</tr>
<tr>
<td>NM_016521</td>
<td>0</td>
<td>NM_007111</td>
<td>21</td>
<td>transcription factor Dp-3</td>
<td>transcription factor Dp-1</td>
<td>2.86E-06</td>
</tr>
<tr>
<td>NM_018677</td>
<td>25</td>
<td>NM_032501</td>
<td>2</td>
<td>ACAS2 (ADP forming)</td>
<td>ACAS2 (AMP forming)-like</td>
<td>7.62E-06</td>
</tr>
<tr>
<td>NM_002293</td>
<td>22</td>
<td>NM_006059</td>
<td>1</td>
<td>laminin, gamma 1</td>
<td>laminin, gamma 3</td>
<td>9.72E-06</td>
</tr>
<tr>
<td>NM_006366</td>
<td>1</td>
<td>NM_006367</td>
<td>22</td>
<td>CAP, adenylate cyclase-associated protein, 2</td>
<td>CAP, adenylate cyclase-associated protein 1</td>
<td>9.72E-06</td>
</tr>
<tr>
<td>NM_005061</td>
<td>0</td>
<td>NM_00967</td>
<td>19</td>
<td>ribosomal protein L3-like</td>
<td>ribosomal protein L3</td>
<td>1.05E-05</td>
</tr>
<tr>
<td>NM_015528</td>
<td>19</td>
<td>NM_181710</td>
<td>0</td>
<td>ring finger protein 167</td>
<td>zinc and ring finger 4</td>
<td>1.05E-05</td>
</tr>
<tr>
<td>NM_001416</td>
<td>29</td>
<td>NM_014740</td>
<td>4</td>
<td>eukaryotic translation initiation factor 4A</td>
<td>DEAD (Asp-Glu-Ala-Asp) box polypeptide 48</td>
<td>1.12E-05</td>
</tr>
<tr>
<td>NM_004739</td>
<td>3</td>
<td>NM_004689</td>
<td>26</td>
<td>MTA2</td>
<td>metastasis associated 1</td>
<td>1.7E-05</td>
</tr>
<tr>
<td>NM_017772</td>
<td>1</td>
<td>NM_014346</td>
<td>21</td>
<td>TBC1, member 22B</td>
<td>TBC1, member 22A</td>
<td>1.79E-05</td>
</tr>
<tr>
<td>NM_00209</td>
<td>21</td>
<td>NM_000887</td>
<td>1</td>
<td>integrin, alpha L</td>
<td>integrin, alpha X</td>
<td>1.79E-05</td>
</tr>
<tr>
<td>NM_000224</td>
<td>21</td>
<td>NM_019010</td>
<td>1</td>
<td>keratin 18</td>
<td>keratin 20</td>
<td>1.79E-05</td>
</tr>
<tr>
<td>NM_004953</td>
<td>37</td>
<td>NM_003760</td>
<td>9</td>
<td>eukaryotic translation initiation factor 4A</td>
<td>eukaryotic translation initiation factor 4 gamma, 3</td>
<td>3.08E-05</td>
</tr>
<tr>
<td>NM_182743</td>
<td>27</td>
<td>NM_000637</td>
<td>4</td>
<td>thioredoxin reductase 1</td>
<td>glutathione reductase</td>
<td>3.31E-05</td>
</tr>
<tr>
<td>NM_003380</td>
<td>31</td>
<td>NM_001927</td>
<td>6</td>
<td>vimentin</td>
<td>desmin</td>
<td>3.51E-05</td>
</tr>
<tr>
<td>NM_021804</td>
<td>0</td>
<td>NM_152831</td>
<td>17</td>
<td>angiotensin I converting enzyme (peptidyl-</td>
<td>angiotensin I converting enzyme (peptidyl-dipeptidase A) 1</td>
<td>3.81E-05</td>
</tr>
<tr>
<td>NM_198597</td>
<td>24</td>
<td>NM_014822</td>
<td>3</td>
<td>SEC24 related gene family, member C</td>
<td>SEC24 related gene family, member D</td>
<td>5.19E-05</td>
</tr>
<tr>
<td>NM_152856</td>
<td>32</td>
<td>NM_005778</td>
<td>7</td>
<td>RNA binding motif protein</td>
<td>RNA binding motif protein 10</td>
<td>5.61E-05</td>
</tr>
<tr>
<td>NM_019854</td>
<td>4</td>
<td>NM_198319</td>
<td>26</td>
<td>HMT1 hnRNP methyltransferase-like 4</td>
<td>HMT1 hnRNP methyltransferase-like 2</td>
<td>5.65E-05</td>
</tr>
<tr>
<td>NM_006431</td>
<td>4</td>
<td>NM_006429</td>
<td>26</td>
<td>chaperonin containing TCP1, subunit 2 (beta)</td>
<td>chaperonin containing TCP1, subunit 7 (eta)</td>
<td>5.65E-05</td>
</tr>
<tr>
<td>NM_006472</td>
<td>19</td>
<td>NM_183376</td>
<td>1</td>
<td>thioredoxin interacting protein arrestin domain containing 4</td>
<td>thioredoxin interacting protein arrestin domain containing 4</td>
<td>6.06E-05</td>
</tr>
<tr>
<td>NM_021971</td>
<td>2</td>
<td>NM_205847</td>
<td>21</td>
<td>GDP-mannose pyrophosphorylase B</td>
<td>GDP-mannose pyrophosphorylase A</td>
<td>7.83E-05</td>
</tr>
</tbody>
</table>
Table 1. Continue

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>AS forms of gene 1</th>
<th>Gene 2</th>
<th>AS forms of gene 2</th>
<th>Function of gene 1</th>
<th>Function of gene 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_012302 21</td>
<td>NM_014921 2</td>
<td>latrophilin 2</td>
<td>latrophilin 1</td>
<td>7.83E-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_054013 21</td>
<td>NM_012214 2</td>
<td>MGAT4B, transferase activity</td>
<td>MGAT4A, transferase activity</td>
<td>7.83E-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_000477 29</td>
<td>NM_001134 6</td>
<td>albumin</td>
<td>α-fetoprotein</td>
<td>9.55E-05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Summary of expression and alternative splicing divergence of duplicates

<table>
<thead>
<tr>
<th>Duplicate pair categories</th>
<th>P&lt;0.01(181)</th>
<th>P&gt;0.01(2694)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of pairs</td>
<td>Proportion</td>
</tr>
<tr>
<td>HH</td>
<td>89</td>
<td>0.49</td>
</tr>
<tr>
<td>LL</td>
<td>9</td>
<td>0.05</td>
</tr>
<tr>
<td>LH</td>
<td>83</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Shown here is the classification of the two groups of duplicate pairs which have different AS form difference level based on EST coverage. HH, both pairs have more than 112 EST hits. LL, both pairs have equal or less than 112 EST hits. LH, one gene of duplicate pair has more than 112 EST hits and another one has equal or less than 112 EST hits.
Figure legends:

Figure 1: The distribution of alternative splice (AS) forms in human genes. (A) Each bar represents the percentage of human genes with the given number of AS forms. (B) The number of genes decays with the number of AS forms, following an exponential law (>19 AS forms excluded for simplicity).

Figure 2: Schematic illustration for the evolution of alternative splicing after gene duplication. The ancestral gene has two alternative splice forms, L (long) and S (short). In the early stage after gene duplication, the L and S forms may become dominantly expressed in one of duplicate copies A and B, respectively. In the late stage, some novel alternative splice forms may be generated.

Figure 3: Less alternative splice forms in larger gene families. Error bar shows the standard error. (A) Mean number of AS forms plotting against human gene family size (size = 1 means single-copy gene). (B) Proportion of genes that are not alternatively spliced plotting against the human gene family size.

Figure 4: The proportion of no alternatively spliced genes (A) and the mean of alternative splice forms (B) plotting against the protein distance between duplicates (with a bin of 0.1 distance unit). Error bar shows the standard error.

Figure 5: Recent duplicated genes are unlikely to be alternatively spliced. Each bar represents the proportion of genes that are not be alternatively spliced in four model organisms. Genes have the sequence identity greater than 30% and E-value less than 1E-10 were classified to the group 30. Similarly, the groups 50, 70 and 90 were under the identity cutoff 50%, 70% and 90% respectively.
Figure 6: The distribution of the number of alternative splice form difference in all duplicate pairs. White and black bars represent the P-value (asymmetric AS evolution between duplicates) less than 0.01 and more than 0.01 respectively.

Figure 7: The P-value distribution of all duplicate pairs for asymmetric AS evolution after gene duplication, which were calculated by binomial test under the null hypothesis of no difference in the number of alternative splice forms (see Methods).
Figure 1:

A

B

y = 3285.8e^{-0.2393x}

R^2 = 0.9972
Figure 2:

Early stage after gene duplication  Late stage after gene duplication
Figure 3:

A

B

Mean of AS forms

Proportion of no AS

Gene family size

Gene family size
Figure 4:

A

Proportion of no AS

B

Mean AS number

Protein Distance

Protein Distance
Figure 5:

A. Proportion of no AS for Human

B. Proportion of no AS for Mouse

C. Proportion of no AS for Drosophila

D. Proportion of no AS for C. elegans

Protein sequence identity cutoff vs. Proportion of no AS
Figure 7:

A

B

Number of pairs

Number of pairs

P-value

P-value
CHAPTER 5 ALTERNATIVE SPLICING AND PROTEIN DIVERSITY IN MAMMALS, INVERTEBRATES AND PLANTS

Jianmin Wang, Liang Ye, Jianjun Chen, Xun Gu, Xiaoqiu Huang

Abstract

Alternative splicing (AS) is considered to be a major source of protein diversity in vertebrates. Here we compare the role of alternative splicing (AS) in the generation of protein diversity in mammals, invertebrates, and plants by assessing the differences in AS rate among six species, two from each class. To address variation in coverage of mRNA sequences by expressed sequence tags (ESTs), we constructed subsets of mRNA sequences from each species under each of several EST coverage conditions. One of the EST coverage conditions was that each mRNA sequence in the subset has at least three exons that are each coverable by three EST sequences. We estimated the AS rate for each subset of mRNA sequences by randomly generating 1,000 subsets of EST sequences with each EST subset satisfying the coverage condition of the mRNA subset. The AS rates for six subsets of mRNA sequences under the EST coverage condition above are 51% (human), 43% (mouse), 34% (fruit fly), 28% (rice), 22% (nematode), and 16% (mustard). The margin of error at the 99% confidence level for each AS rate is below 1%. We therefore conclude that alternative splicing also plays a significant role in invertebrates and plants compared to mammals.

Introduction

Alternative splicing (AS) is considered to play a more significant role in the generation of protein diversity in vertebrates than in invertebrates based on the unexpectedly small difference in gene number in fruit flies (Drosophila), nematodes (C. elegans), mice, and humans (Black 2000; Graveley 2001). However, it is a conceptual and practical challenge to
assess this proposal by using existing mRNA and expressed sequence tag (EST) data because of various limitations of the data (Modrek et al. 2001; Brett et al. 2002; Kapranov et al. 2002; Kim et al. 2004; Harrington et al. 2004). Because some plants have more genes than animals, AS may be significantly less prevalent in mustards (*Arabidopsis*) and rices than in humans (Kikuchi et al. 2003; Iida et al. 2004). However, it is suggested that AS is more prevalent in plants than was previously thought (Kazan 2003).

The AS rate of a species can be estimated by matching EST sequences to mRNA sequences called reference sequences (RefSeqs) from the species (Hanke et al. 1999; Mironov et al. 1999; Croft et al. 2000; Kan et al. 2001; Modrek et al. 2001; Modrek and Lee 2002). The estimate increases with EST coverage (Brett et al. 2000; Kan et al. 2001; Brett et al. 2002). In addition, the estimate varies with the number of exons of the gene (Johnson et al. 2003). Another issue is that some EST sequences are problematic, leading to a false alternatively spliced (AS) form (Modrek et al. 2001). Those issues have resulted in inconsistent assessments of differences in AS rate among a number of species (Brett et al. 2002; Kim et al. 2004).

Here we compare the extent to which AS contributes to protein diversity in mammals, invertebrates, and plants by assessing the differences in AS rate in representative species from each class: human, mouse, fruit fly, nematode, rice, and mustard. We develop a statistical sampling method for using incomplete mRNA and EST data to assess the differences in AS rate among the six species. The method is designed to address the issues in estimation of the AS rates in the six species. In the method, subsets of RefSeq sequences for the six species are constructed under various EST coverage conditions. For each family of six RefSeq subsets under the same EST coverage condition, the AS rate for each RefSeq subset
is estimated by randomly generating 1,000 subsets of EST sequences with each EST subset satisfying the coverage condition of the RefSeq subset. In addition, the random generation of subsets of EST sequences reduces the effect of problematic ESTs on the AS rate.

We constructed subsets of RefSeq sequences for the six species under each of several EST coverage conditions and estimated the AS rate for each subset of RefSeq sequences by randomly generating 1,000 subsets of EST sequences with each EST subset covering the RefSeq subset under the same coverage condition. Our results show that AS is prevalent in the three classes of organisms: mammals, invertebrates, and plants. At every EST coverage depth, the AS rate for the human subset of RefSeq sequences is slightly higher than that for the mouse subset. The AS rate for every rice subset is about twice as high as that for the corresponding mustard subset. At every EST coverage depth, human and mouse subsets have notably higher AS rates than the subsets for the other four species.

Results

For each of the six species, its RefSeq and EST sequences were mapped to its genome. For each species, the numbers of RefSeq and EST sequences before and after the mapping are shown in Table 1. For each genomic locus with mapped RefSeq sequences, a mapped RefSeq sequence with a longest match to the locus was selected for identification of AS events. An exon of a selected RefSeq sequence was covered by an EST sequence if the EST sequence was mapped to the same locus as the RefSeq sequence and the match between the exon and the EST sequence either covered at least 50% of the exon or was at least 100 bases long. Two source sets were constructed, one source set with all selected RefSeq sequences and the other set with all EST sequences that cover an exon of a selected RefSeq sequence. See Methods for details.
Subsets of RefSeq sequences were constructed under an EST coverage condition from the source sets of RefSeq sequences for the six species such that the subsets are comparable in EST coverage. The EST coverage condition for a RefSeq sequence is characterized by using two numbers $m$ and $n$, where $m$ is the number of selected exons of the RefSeq sequence and each selected exon can be covered by at least $n$ EST sequences. Each combination $(m, n)$ represents a family of six subsets of RefSeq sequences, one subset per species. For example, combination $(2, 3)$ represents the family of six subsets of RefSeq sequences that were constructed as follows. For each of the six species, a sequence from the source set of RefSeq sequences for the species was included in the subset for the species in combination $(2, 3)$ if the sequence has at least 2 exons such that each exon can be covered by at least 3 EST sequences from the source set of EST sequences for the species.

Table 2 shows, for each species and for each of six combinations $(m, n)$ with $1 \leq m \leq 3$ and $n = 1$ or $3$, the size of the RefSeq subset for the species in the combination. Note that the RefSeq subsets for fruit fly, nematode, rice, and mustard are much smaller than the RefSeq subsets for human and mouse for some combinations in Table 2. For example, the smallest subset for combination $(3, 3)$ contains 3434 RefSeq sequences from rice. To see variation in AS rate, we considered nine combinations $(m, n)$ with $1 \leq m \leq 3$ and $1 \leq n \leq 3$ for assessing the AS rate for each subset of RefSeq sequences in each of the combinations. Because of the small sizes of the subsets of RefSeq sequences for fruit fly, nematode, rice, and mustard in combinations $(m, n)$ with $m > 3$ or $n > 3$, we considered only the nine combinations. The sizes of the RefSeq subsets in combinations $(m, n)$ with $1 \leq m \leq 20$ and $1 \leq n \leq 50$ are provided in Supplementary Material.
For each combination \((m, n)\) and each subset of RefSeq sequences in the combination, the AS rate for the RefSeq subset at EST coverage depth \(n\) was estimated by randomly generating 1,000 subsets of EST sequences with each subset of EST sequences covering the subset of RefSeq sequences to coverage depth \(n\). A subset of EST sequences that covers the subset of RefSeq sequences to coverage depth \(n\) was randomly generated as follows. Recall that each sequence in the RefSeq subset has at least \(m\) exons such that each exon can be covered by at least \(n\) EST sequences in the source set. For each RefSeq sequence in the subset, \(m\) different exons with each coverable by EST sequences to depth \(n\) were randomly selected and for each of the \(m\) exons, \(n\) EST sequences that cover the exon were randomly selected from the EST source set and were added to the EST subset. In other words, for each subset of EST sequences, each RefSeq sequence in the subset has \(m\) randomly selected exons with each exon covered by \(n\) EST sequences in the subset.

For each of the 1,000 subsets of EST sequences, an AS rate for the RefSeq subset was computed by using only EST sequences in the EST subset. The mean of the 1,000 AS rates along with a 99% confidence interval for the RefSeq subset was computed. The 99% confidence interval was determined such that 990 of the 1,000 AS rates fell into the confidence interval. Figure 1 shows the mean AS rate and its 99% confidence interval for each subset of RefSeq sequences at EST coverage depth \(n\) in each of the 9 combinations \((m,n)\).

The results in Figure 1 show that the AS rate for each human RefSeq subset is slightly higher than that for the mouse RefSeq subset in every combination. In every combination, the AS rate for the rice subset of RefSeq sequences is about twice as high as the AS rate for the mustard subset. In addition, in every combination, both human and mouse subsets of RefSeq
sequences have notably higher AS rates than the subsets of RefSeq sequences for fruit fly, nematode, rice, and mustard. In combination (3, 3), both fruit fly and rice subsets of RefSeq sequences have sufficiently high AS rates.

There is another variable in addition to the variables $m$ and $n$ for the results in Figure 1. The variable is the subset of RefSeq sequences. The subsets of RefSeq sequences for the same species in different combinations of Figure 1 are different in size. To eliminate the effect of the variable on the AS rate, we fixed the subset of RefSeq sequences for each species to that in combination (1, 5) and computed the AS rate for each subset of RefSeq sequences at each EST coverage depth $d$ from 1 to 5. For each subset of RefSeq sequences and each EST coverage depth $d$ from 1 to 5, 1,000 subsets of EST sequences with each covering the subset of RefSeq sequences at coverage depth $d$ were randomly generated to compute the mean AS rate along with its 99% confidence interval for the subset of RefSeq sequences. The results in Figure 2 show that the AS rates for all the six subsets of RefSeq sequences increase significantly with EST coverage depth $d$. In addition, the AS rates for human and mouse subsets of RefSeq sequences increase with EST coverage depth at a faster pace than the rates for the other four subsets of RefSeq sequences.

Because there are sufficiently large human and mouse RefSeq subsets under a high EST coverage condition, we estimated the AS rates for the human and mouse RefSeq subsets in combination (5, 20). The sizes of the human and mouse RefSeq subsets are 4838 and 4031, respectively, whereas the maximum size of the RefSeq subsets for the other four species in the combination is 222. The mean AS rate for the human RefSeq subset at EST coverage depth 20 is 84.9% and its 99% confidence interval is [84.0%, 85.7%], whereas the mean AS rate for the mouse RefSeq subset is 76.9% and its 99% confidence interval is [76.0%, 77.9%].
Discussion

Because of low EST coverage for species other than human and mouse, it is difficult to estimate the real AS rates for those species. However, creating subsets of RefSeq sequences under a low EST coverage condition for the six species allowed us to assess and compare the AS rates for the subsets under the same EST coverage condition. Our results show that the AS rate for each human RefSeq subset is slightly higher than that for the mouse RefSeq subset under each of EST coverage conditions. There are notable differences in AS rate between human/mouse RefSeq subsets and RefSeq subsets from fruit fly, nematode, rice and mustard under low EST coverage conditions. The AS rate for each rice RefSeq subset is about twice as high as that for the corresponding mustard subset. The statistically significant differences in AS rate among RefSeq subsets from those species under low EST coverage conditions indicate that alternative splicing contributes to the generation of protein diversity at different levels in those species.

We considered a number of low EST coverage conditions under which sufficiently large RefSeq subsets could be constructed. A high EST coverage condition would lead to higher AS rates for the subsets of RefSeq sequences under the EST coverage condition than a low EST coverage condition. On the other hand, the high EST coverage condition will lead to much smaller subsets of RefSeq sequences than the low EST coverage condition. Thus, an AS rate obtained by using a smaller subset of RefSeq sequences from a species may be further away from the AS rate that would be obtained by using the full set of RefSeq sequences under the EST coverage condition if the full set was available.

Our results show that the AS rates for RefSeq subsets are very sensitive to the EST coverage condition, but are not sensitive to differences among EST subsets that meet the
coverage condition. The differences among EST subsets include different EST sequence quality, different EST tissue sources, and different exons covered by ESTs.

The high AS rates for human, mouse, fruit fly, and rice RefSeq subsets under low EST coverage conditions suggest that those species have very high AS rates. We therefore conclude that alternative splicing plays a significant role in generation of protein diversity in invertebrates, vertebrates, and plants.

**Methods**

We first provide the databases for all sequence data used in this study. Then we describe a method for constructing a source set of RefSeq sequences and a source set of EST sequences for each species by mapping its RefSeq and EST sequences to its genome sequences. Next we present a method for finding AS forms in a subset of RefSeq sequences by using a subset of EST sequences. Finally we describe a sampling method for constructing subsets of RefSeq sequences under various EST coverage conditions and calculating the AS rate for each RefSeq subset by randomly generating 1,000 subsets of EST sequences with each EST subset meeting the coverage condition.

**Sequence data**

The RefSeq sequences for human, mouse, fruit fly, nematode, and mustard were downloaded from the RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/refseq). The full-length cDNA sequences for rice were downloaded from (http://cdna01.dna.affrc.go.jp/cDNA) and were referred to as the RefSeq sequences for rice throughout the paper. The EST sequences for all the six species were downloaded from the dbEST database (ftp://ftp.ncbi.nlm.nih.gov/repository/dbEST). The genome sequences of human, mouse, fruit fly and nematode were downloaded from the UCSC genome database.
The genome sequence of rice was downloaded from TIGR (http://www.tigr.org). The genome sequence of mustard was downloaded from TAIR (http://www.arabidopsis.org).

**Construction of source sets of RefSeq and EST sequences**

The RefSeq and EST sequences for each species were mapped to the genome sequence of the species by the BLAT program (Kent, 2002). A match between a RefSeq/EST sequence and a genomic region is acceptable if the match covers at least 75% of the RefSeq/EST sequence and the percent identity of the match is at least 95%. If a RefSeq/EST sequence is matched to multiple genomic regions, then the RefSeq/EST sequence is assigned to the genomic region with the best match. If two genomic regions have the same best match with the RefSeq/EST sequence, then the RefSeq/EST sequence is discarded. For each match between an EST sequence and a genomic region, candidate exons and introns in the genomic region are found from the match. If candidate exons or introns are extremely short, then the EST sequence is discarded. If the boundary between adjacent candidate exon and intron contains no significant splice site signal from the GeneSplicer program (Pertea et al. 2001), then the EST sequence is discarded.

If two or more RefSeq sequences are mapped to the same genome region, then a RefSeq sequence with the longest genome coverage is selected. For any genomic region, at most one RefSeq sequence can be selected for the region, whereas multiple EST sequences can be assigned to the region. The EST sequences assigned to a region are associated with the selected RefSeq sequence for the region. A source set of selected RefSeq sequences and a source set of assigned EST sequences were constructed for identification of AS forms for the species.
**Detection of AS forms**

For each selected RefSeq sequence and each of the EST sequences associated with the RefSeq sequence, the match between the sequence and the genomic region is used to find the exons in the sequence. The pattern of splicing for the sequence is represented by a list of its exons in order of increasing coordinates. Alternatively spliced forms for the RefSeq sequence are found by comparing the exon list of each EST sequence with the exon list of the RefSeq sequence and finding differences in exon coordinates (Su et al. 2006). Redundant AS forms were removed.

For a subset of the RefSeq source set and a subset of the EST source set for a species, the AS rate for the RefSeq subset with respect to the EST subset is computed as follows. For each RefSeq sequence in the subset, AS forms for the sequence are identified by using only the EST sequences in the subset. The AS rate for the RefSeq subset with respect to the EST subset is the number of RefSeq sequences with an AS form divided by the total number of RefSeq sequences in the subset.

**Construction of subsets of RefSeq and EST sequences**

An EST coverage condition for constructing a subset of RefSeq sequences is specified by using two numbers \( m \) and \( n \), where every RefSeq sequence in the subset has at least \( m \) exons such that each of the \( m \) exons can be covered by at least \( n \) EST sequences. A subset of RefSeq sequences is constructed under an EST coverage condition \((m, n)\) as follows. Every RefSeq sequence in the source set of RefSeq sequences is examined. If the RefSeq sequence has at least \( m \) exons, each of which can be covered by at least \( n \) EST sequences, then the RefSeq sequence is included in the subset. For the subset of RefSeq sequences under the EST coverage condition \((m, n)\), a subset of EST sequences that meets the coverage condition is
randomly constructed as follows. For each RefSeq sequence in the subset, $m$ exons of the RefSeq sequence with each exon coverable by $n$ EST sequences are randomly selected. For each selected exon, $n$ EST sequences that cover the exon are randomly selected, and are included in the subset of EST sequences.

The AS rate for a subset of RefSeq sequences is estimated by randomly generating 1,000 subsets of EST sequences with each EST subset satisfying the EST coverage condition of the RefSeq subset. For each EST subset, the AS rate of the RefSeq subset is computed with respect to the EST subset. The mean AS rate along with a 99% confidence interval for the RefSeq subset is computed by using the 1,000 AS rates. The 99% confidence interval is determined such that 99% of the 1,000 AS rates are inside the interval.

Acknowledgments

This work was supported in part by NIH grants.

References


Figure legends

**Figure 1:** Estimations of AS rates for the six RefSeq subsets in each of 9 combinations \((m, n)\) with \(1 \leq m \leq 3\) and \(1 \leq n \leq 3\). For each combination \((m, n)\), the bar and error bar show the mean AS rate and its 99% confidence interval for each of the six RefSeq subsets at EST coverage depth \(m\) in the combination. The six RefSeq subsets in left-right order are from human, mouse, fruit fly, nematode, rice, and mustard, respectively.

**Figure 2:** The AS rates for the six RefSeq subsets in combination \((1, 5)\) under each EST coverage depth from 1 to 5.
Figure 1:
Figure 2:

![Graph showing AS rate vs. depth of coverage for different species.](image)

- **human**
- **mouse**
- **drosophila**
- **c. elegans**
- **rice**
- **arabidopsis**
Tables

Table 1. Number of RefSeq and EST sequences before and after genome mapping

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of Refseqs</th>
<th>No. of ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>Human</td>
<td>21538</td>
<td>21369</td>
</tr>
<tr>
<td>Mouse</td>
<td>17213</td>
<td>17062</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>18786</td>
<td>18695</td>
</tr>
<tr>
<td>Nematode</td>
<td>21121</td>
<td>20792</td>
</tr>
<tr>
<td>Rice</td>
<td>32123</td>
<td>29739</td>
</tr>
<tr>
<td>Mustard</td>
<td>28953</td>
<td>28818</td>
</tr>
</tbody>
</table>

Table 2. The sizes of the subsets of RefSeq sequences in six combinations

<table>
<thead>
<tr>
<th>Species</th>
<th>(1,1)</th>
<th>(1,3)</th>
<th>(2,1)</th>
<th>(2,3)</th>
<th>(3,1)</th>
<th>(3,3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>15276</td>
<td>14689</td>
<td>14291</td>
<td>13312</td>
<td>13232</td>
<td>12172</td>
</tr>
<tr>
<td>Mouse</td>
<td>14094</td>
<td>13371</td>
<td>12904</td>
<td>12062</td>
<td>11917</td>
<td>11017</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>9461</td>
<td>7808</td>
<td>7451</td>
<td>5696</td>
<td>5281</td>
<td>3663</td>
</tr>
<tr>
<td>Nematode</td>
<td>14328</td>
<td>9108</td>
<td>13650</td>
<td>7949</td>
<td>11877</td>
<td>6377</td>
</tr>
<tr>
<td>Rice</td>
<td>13339</td>
<td>8024</td>
<td>9297</td>
<td>4819</td>
<td>7062</td>
<td>3434</td>
</tr>
<tr>
<td>Mustard</td>
<td>16803</td>
<td>12215</td>
<td>13052</td>
<td>7354</td>
<td>10138</td>
<td>4756</td>
</tr>
</tbody>
</table>
CHAPTER 6 GENERAL CONCLUSION

In this dissertation, we first described an assembly program for sequences with deep coverage. Then we developed a program to identify SNPs and compute allele frequencies using sequences from individuals. At last, we developed a pipeline to identify alternative splicing forms. We used the pipeline to study the relationship between alternative splicing and gene duplication; we also used the pipeline to compare alternative splicing among six species and explored the diversity of alternative splicing in plants.

In Chapter 2, we described a deep assembly program named DA. The program was designed to handle sequences with deep coverage, which are common for lots of highly expressed genes. The major computation difficulties are that the number of alignments to be computed is huge and that the sequence error rates in ESTs are high. Several strategies are applied in the program so that it can only compute a linear number of alignments relative to the number of input sequences. Experimental results on the 105 clusters from the UniGene database show that DA can handle data sets of deep coverage efficiently. Also DA can generate accurate consensus sequences for the test datasets by comparing the results with known genomic sequences for human and mouse.

Even though DA can handle sequences with deep coverage well, it can not fully identify alternative splice forms, which are common for those sequences. DA can only identify some of the alternative splice patterns and does not have the ability to generate all transcripts correctly. Two possible approaches can be used to deal with the problem. In the first approach, each input sequence can be used multiple times, so it can appear in several alternative transcripts. This method can generate multiple transcripts and all of them are compatible with the input sequences if a good clipping method is applied. But it may miss
some real transcripts and generate some wrong variants, because EST sequences are from only partial transcripts. Another approach is building a splice graph from EST sequences. In the splice graph, each node is a contig with partial or complete multiple input sequences and each edge represents a splice pattern. This method can not generate any complete transcript; further analysis is needed to get possible transcripts. The second strategy is more flexible and can provide more information to study alternative splicing.

SNPs are important genetic markers that can be used in association studies. EST sequences can provide a cost-inexpensive way to discover SNPs. In Chapter 3, we described a computational method for finding common SNPs with allele frequencies in single-pass sequences of deep coverage. The method focuses on distinguishing sequencing errors from true SNPs. The method also tries to assign each individual sequence to the true genomic position instead of paralogous positions. The method enhances a widely used program named PolyBayes in several aspects. The results from our method and PolyBayes on eighteen data sets of human expressed sequence tags (ESTs) with deep coverage indicate that our method used almost all single-pass sequences in computation of the allele frequencies of SNPs.

The method of distinguishing sequence errors and true SNPs (including heterogeneous bases) and the method of identifying paralogous sequences for sequences with deep coverage can be improved by using good statistical models too. PolyBayes provides good models for low coverage case and it does not consider heterogeneous bases. Those can be improved. Also for EST sequences, library information can be used to improve the accuracy of estimating allele frequencies. For any diploid individual or library, only two chromosomes will contribute in the count of allele frequencies.
EST sequences can also be used to study alternative splicing (AS), which is the most common post transcription event in metazoans. We developed a pipeline to identify AS forms by comparing alignments between expressed sequences and genomic sequences. In Chapter 5, we studied the relationship between AS and gene duplication. We observed that duplicate genes have fewer AS forms than single-copy genes; we also found that the loss of alternative splicing in duplicate genes may occur shortly after gene duplication. These results support the subfunctionization model of alternative splicing in the early stage after gene duplication. Further analysis of the alternative splicing distribution in human duplicate pairs showed the asymmetric evolution of alternative splicing after gene duplication. The above study used sequences from human and mouse. Further studies on other species can provide even more insights about the relationship between alternative splicing and gene/genome duplication. Plants have both ancient and recent genome duplications, which provide opportunities to study the topic too.

In Chapter 6, we also compared AS among six species. We found significant differences on both AS rates and splice forms per gene among the species through detailed and categorized studies. We discovered that the AS rate in Arabidopsis is similar to that in previous studies, but rice has a much higher AS rate than that of previous studies. The difference in AS between rice and Arabidopsis is significant, so the difference in protein diversity among plants could be tremendous. In the study, we already found that the distribution of AS rates among species is continuous. The differences between species can be highly significant, like human and Arabidopsis, but it can also be small, like human and mouse. With more species being included in the study, a more complete view of the distribution can be acquired and it can be imagined that the change could be pretty smooth.
To better estimate and compare the true AS rates, EST coverage for all genes should be high enough. Some alternative splice forms are tissue, developmental stage and environment dependent and some are related to disease. So the study of AS under different situations can be used to answer biological questions.