Abstract

YOUNG, STUART. Improved EST Annotation Using Multiple External Databases. (Under the direction of Douglas L. Crawford.)

Expressed sequence tags (ESTs) are the largest existing genomic resource and play an important role in a wide range of genomic research areas. Correct and useful functional annotation of EST resources enhances their use with other genomic resources.

Chapter 2 describes problems with the use of BLAST sequence similarity searches to annotate new species-specific sequence databases. To reduce annotation errors, an incremental approach is proposed based on comprehensive database searches to verify functionally similar BLAST hits. In Chapter 3, machine learning models based on protein motifs are proposed as an alternative or adjunct to sequence similarity methods for functional annotation. This motif-based approach is particularly suited to the annotation of sequence fragments such as EST collections. In Chapter 4, a methodology to improve EST collection annotation by a series of quality control steps is described. This methodology was implemented on the updated *F. heteroclitus* EST database using myEST, an open-source EST annotation platform. The myEST platform implements the methods discussed in the previous chapters and also incorporates additional steps integrating data from multiple external databases to improve EST quality control, provide more accurate annotation and facilitate interactive sequence analysis.
Improved EST Annotation Using Multiple External Databases

by

Stuart Young

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy Bioinformatics Raleigh, North Carolina 2008

APPROVED BY:

__________________________         _________________________
Gregory C. Gibson            Craig Sullivan
Chair of Advisory Committee

__________________________         _________________________
Spencer Muse                Steffen Heber

__________________________
Douglas Crawford
Dedication

To Yuching, Rio and Shaw and my loving family.
Biography

Stuart Young earned his B.Sc. Biological Sciences from the University of Leicester, England, in 1989 before embarking on a career in journalism. Returning to his original interest in the life sciences, he graduated with an M.S. Bioinformatics from Indiana University Bloomington in 2005.
# Table of Contents

List of Tables .................................................................................................................. vii

List of Figures .................................................................................................................. ix

Introduction ....................................................................................................................... 1

1. Annotation ..................................................................................................................... 1
   1.1 Homology .................................................................................................................. 2
   1.2 Analogy ..................................................................................................................... 2
   1.3 Orthology .................................................................................................................. 2
   1.4 Paralogy .................................................................................................................... 3
   1.5 Sequence similarity ................................................................................................. 3

2. BLAST E-value Cutoffs ................................................................................................. 4

3. Expressed Sequence Tags (ESTs) .................................................................................. 5
   3.1 Condition-specific expression .................................................................................. 6
   3.2 Gene identification in genomic data ........................................................................ 8
   3.3 Alternative splicing prediction ................................................................................. 10
   3.4 Caveat: A High Proportion of Uncharacterized ESTs ............................................. 11

4. A short history of *F. heteroclitus* ............................................................................ 11
   4.1 Phylogeny ................................................................................................................. 12
   4.2 *Fundulus heteroclitus* as a Genomic Resource ..................................................... 13

5 Machine learning using protein motif inputs .............................................................. 18
   5.1 An introduction to Decision Trees .......................................................................... 19
   5.2 An introduction to Naïve Bayes .............................................................................. 20

References ....................................................................................................................... 23

Chapter 2: Incremental mismatch probability filter improves BLAST annotation accuracy .... 34

Abstract ............................................................................................................................ 35

Introduction ....................................................................................................................... 36

Methods .............................................................................................................................. 38

   2.1 UniGene and Swiss-Prot BLAST Comparisons ...................................................... 38
   2.2 BLAST scores, Histograms and E-value ................................................................. 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3</td>
<td>Fuzzy Matching of BLAST Annotations</td>
<td>39</td>
</tr>
<tr>
<td>2.4</td>
<td>Incremental mismatch probability to validate BLAST hits</td>
<td>40</td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>3.1</td>
<td>Estimated proportion of missing orthologues</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>Estimation of annotation quality</td>
<td>43</td>
</tr>
<tr>
<td>3.3</td>
<td>Estimating the similarity of E-values</td>
<td>44</td>
</tr>
<tr>
<td>3.4</td>
<td>Fuzzy Match Algorithm matches similar annotations</td>
<td>45</td>
</tr>
<tr>
<td>3.5</td>
<td>The utility of the incremental mismatch probability statistic</td>
<td>48</td>
</tr>
<tr>
<td>Conclusion</td>
<td></td>
<td>54</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Chapter 3: InterPro Motif Annotation Trees (IMAT)</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td></td>
<td>74</td>
</tr>
<tr>
<td>Introduction</td>
<td></td>
<td>75</td>
</tr>
<tr>
<td>Methods</td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>2.1</td>
<td>Motif distributions and correlations</td>
<td>77</td>
</tr>
<tr>
<td>2.2</td>
<td>Validating Decision Trees</td>
<td>77</td>
</tr>
<tr>
<td>2.3</td>
<td>Human UniGene orthologues</td>
<td>78</td>
</tr>
<tr>
<td>2.4</td>
<td>CluSTr orthologues</td>
<td>78</td>
</tr>
<tr>
<td>2.5</td>
<td>Eukaryotic orthologues (KOG)</td>
<td>79</td>
</tr>
<tr>
<td>2.6</td>
<td>Orthologue training/test set selection</td>
<td>79</td>
</tr>
<tr>
<td>2.7</td>
<td>Swiss-Prot tree training/test set selection</td>
<td>80</td>
</tr>
<tr>
<td>2.8</td>
<td>Motif categorization</td>
<td>81</td>
</tr>
<tr>
<td>2.9</td>
<td>Swiss-Prot Decision Tree evaluation</td>
<td>82</td>
</tr>
<tr>
<td>2.10</td>
<td>InterProScan for motif finding</td>
<td>83</td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td>84</td>
</tr>
<tr>
<td>3.1</td>
<td>Motif distributions and correlations</td>
<td>84</td>
</tr>
<tr>
<td>3.2</td>
<td>Decision Trees outperform Naïve Bayes</td>
<td>85</td>
</tr>
<tr>
<td>3.3</td>
<td>Large, fixed length attributes are optimal</td>
<td>85</td>
</tr>
<tr>
<td>3.4</td>
<td>Swiss-Prot trees predict function</td>
<td>86</td>
</tr>
<tr>
<td>Conclusion</td>
<td></td>
<td>87</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>91</td>
</tr>
</tbody>
</table>
List of Tables

Table 1 - 1 The advantages and disadvantages of Decision Trees and Naive Bayes. ..............29

Table 2 - 1 BLAST of 1,000 human UniGenes. .................................................................62
Table 2 - 2 Number of annotations that are dependent or poor quality...............................63
Table 2 - 3 Relative E-values of the next best hits for three mammalian UniGenes. ..........64
Table 2 - 4 Evaluation of FuzzyMatch performance...........................................................65
Table 2 - 5 Comparison of Swiss-Prot and Mouse Annotations.......................................66
Table 2 - 6 Evaluation of the incremental mismatch probability method to validate 'best hit'
annotations. .......................................................................................................................67

Table 3 - 1 Evaluation of fuzzy matching for human 1,000 and Swiss-Prot annotations. .....94
Table 3 - 2 Pairwise correlations of protein motifs from different InterPro databases among
three orthologue databases and an expressed sequence tags (EST) collection....................95
Table 3 - 3 Pairwise correlations of protein motifs from different InterPro databases for all
Swiss-Prot sequences with InterPro hits (total 276,372). ................................................96
Table 3 - 4 Decision Tree and Naive Bayes orthologue prediction performances for five
different datasets, including UniGene, Clustr and KOG. ...............................................97
Table 3 - 5 Incorrect BLAST hits for 1,000 randomly selected human UniGene against
Swiss-Prot and the mouse, dog and rat UniGene databases. .............................................98
Table 3 - 6 Swiss-Prot protein Decision Tree lookup and prediction performance.............99

Table 4 - 1 Funnybase 2.0 EST tissue sample experiments and read counts. ......................133
Table 4 - 2 The three stages of quality control in the myEST pipeline. The pipeline consists
of a Pre-Assembly, Assembly and Post-Assembly stages which involve different types of
quality control and error checking measures. .................................................................134
Table 4 - 3 Expressed sequence tag (EST) reads from A. californica were successively and
cumulatively assembled using CAP3. The number of contigs and singlets increased with the
number of experiments making up the pool of pre-assembly EST reads, as did the number of
assembled sequences (contigs and singlets) that had BLAST hits against Swiss-Prot and the
RefSeq human, mouse, dog and rat databases (BLAST E-value cutoff: 10^-5). .................135
Table 4 - 4 CAP3 cutoff parameters used in the optimization of K. brevis EST assembly...136
Table 4 - 5 Pre- and post-filter contig and singlet counts for K. brevis EST assemblies using
CAP3 overlap percent identity and overlap length cutoffs...............................................137
Table 4 - 6 The number and quality of BLAST hits and sequence complexity improves
dramatically with pre-assembly filtering of K. brevis sequence reads...............................138
Table 4 - 7 Lack of consistency in Refseq annotation format hinders automated parsing of
synonyms. .......................................................................................................................139
Table 4 - 8 Gene Ontology (GO) mapping files to external databases. ......................140
Table 4 - 9 Synonyms for ‘Putative 2Fe-2S ferredoxin’ derived from the external2go table ordered by relevance.
List of Figures

Figure 1 - 1 Orthologues and paralogues are two types of homologous sequences. ..........30
Figure 1 - 2 Traditional assembly methods often give incorrect results because they fail to
consider alternative splicing. (Pevzner Group).................................................................31
Figure 1 - 3 Distribution of genome sequencing, major EST projects (>1000 sequences),
large-insert libraries and radiation hybrid panels in relation to a conventional fish phylogeny.
.............................................................................................................................................32
Figure 1 - 4 Decision Tree for deciding whether to play golf. .........................................33

Figure 2 - 1 Number of significant hits for 1,000 human UniGenes. .................................68
Figure 2 - 2 FuzzyMatch Algorithm..................................................................................69
Figure 2 - 3 Matched and Mismatched Annotations..........................................................70
Figure 2 - 4 Self-BLAST of Swiss-Prot. ..............................................................................71
Figure 2 - 5 Incremental mismatch probability for human UniGene S15967514 (Cytochrome
c oxidase subunit 8A). ..........................................................................................................72

Figure 3 - 1 Swiss-Prot training/test set selection based on InterPro hits of orthologue set. 100
Figure 3 - 2 Generation of InterPro-derived attributes for input into machine learning
algorithm ...................................................................................................................................101
Figure 3 - 3 Decision Tree generation using InterPro attributes.........................................102
Figure 3 - 4 A workflow for comparing Naive Bayes and Decision Tree models of motif
signatures in three orthologue databases and a 1,000-sequence sample of the human UniGene
EST collection .........................................................................................................................103
Figure 3 - 5 Most InterPro motifs occur in two or more functionally different Swiss-Prot
proteins ......................................................................................................................................104

Figure 4 - 1 Distribution of sequences per organism in the dbEST public EST database. ...142
Figure 4 - 2 The three EST species collections used in the development and testing of the
myEST assembly and annotation platform: Fundulus heteroclitus (mummichog), Karenia
brevis (red tide) and Aplysia californica (sea hare). ..............................................................143
Figure 4 - 3 The synchronous interaction pattern of a traditional web application (top)
compared with the asynchronous communication between server and client of an AJAX
application (bottom) (Garrett 2005). .........................................................................................144
Figure 4 - 4 The myEST plate display function shows marker and cloned sequence
information for each 96-well plate in the collection ..................................................................145
Figure 4 - 5 Cumulative assembly of A. californica sequence read libraries ......................146
Figure 4 - 6 Display of database statistics for the F. heteroclitus EST collection ..................147
Figure 4 - 7 Display of phred quality value distribution of contigs with no BLAST hits .......148
Figure 4 - 8 Two simple metrics for measuring the degree of coverage of contigs predicted from clusters of multiple aligned reads. ................................................................. 149
Figure 4 - 9 Model-View-Controller implementation of the Funnybase 2.0 web interface. 150
Figure 4 - 10 View of the myEST workflow page with run, workflow, stage and application displays. ................................................................. 151
Introduction

The principal focus of this research is the development of improved annotation for expressed sequence tag (EST) databases. We have accomplished this by developing quality control methods for the EST database construction process and by the careful integration of external database searches in the annotation process. This includes a filtering mechanism for BLAST searches which looks at the overall similarity of possible homologues to proteins in a large database, rather than just focusing on the query-target pair, and the use of protein motifs and machine learning techniques to identify unique motif signatures in orthologue groups.

In this introduction we will first briefly clarify what is meant by the term annotation, then discuss in some detail the benefits and pitfalls of EST sequencing and assembly. Finally, we will review the use of the estuarine fish Fundulus heteroclitus as a non-model organism and give a brief overview of the use of machine learning techniques in analyzing biological annotations.

1. Annotation

The term ‘annotation’ – originally a comment, summary or explanatory notes by an author or scholar – has come to mean in the biological sense one or more of the following: protein function, 3-D structure, functional domains, cellular location, post-transcriptional modification, chromosomal location, association with disease traits, promoter elements, evolutionary distance, phenotypic data. In a more abstract sense, annotation is the modeling
of biological objects using terms from predefined vocabularies. The value of the annotation lies in the relevance and usefulness of these vocabularies in defining what it is we need to know about those objects.

The vast majority of biological sequence annotation is based on the notion of homology. To aid further discussion on this topic, we will first define the meanings of some common terms used to describe sequence similarity.

1.1 Homology

Homologous genes are two or more genes in the same or in different species that are derived from a common ancestor (NCBI 2007) (and may or may not share the same function – see Orthology and Paralogy below). Generally, sequence similarity is assumed to correspond to homology. This is under the assumption that sequences are conserved due to evolutionary pressures which act to constrain their structure in order to maintain their function.

1.2 Analogy

Two genes are analogous if they share the same function but are not derived from the same ancestral gene. This is an example of convergent evolution, similar to bird wings and bat wings (Webber and Barton 2003).

1.3 Orthology

In the strict sense, orthology is only concerned with evolutionary events, not function (Figure 1 -1). Orthologues are two or more genes in different species that share the same common
ancestor. By default, orthologues are homologues. Although orthology is defined with respect to evolution and not function, the more recently two genes shared a common ancestor, the more likely it is that they have retained similar functions. Moreover, orthologous genes that have been spared by natural selection from deletion or duplication over many millions of years are also likely to share overlapping functions. The longer two orthologous genes have remained in two different organisms under the pressures of natural selection, the more likely it is that they share the same biological function(s). However, any proposal to classify two orthologous genes as homologues will be strengthened by additional evidence such as shared secondary structure (helices, loops, etc.), tertiary/3D structure and binding site locations.

1.4 Paralogy

Similar to orthology, paralogy is only concerned with evolutionary events, not function (Figure 1-1). Paralogous genes are separated by a gene duplication event. Paralogous genes are often found within a species, but can also be present across species (e.g., the human hemoglobin gene and the chimpanzee myoglobin gene).

1.5 Sequence similarity

Often measured by percent identity, sequence similarity between two sequences is not necessarily the same as orthology or paralogy, but it is a good indicator. Generally speaking, the higher the percentage identity (proportion of residues in common) between two sequences the more likely they are derived from the same common ancestor gene. Support
for the idea of ‘sequence similarity = evolutionary distance = functional distance’ can be found in the classification of isozymes - different genes whose products catalyze the same reaction. In many cases, isozymes are coded for by homologous genes that have diverged over time and the degree of similarity in function is commensurate with the percent sequence identity.

2. BLAST E-value Cutoffs

A reliance on Basic Local Alignment Search Tool (BLAST) (Altschul, et al. 1990) and other sequence alignment tools to identify sequences with some degree of homology is common practice for high-throughput annotation of large biological sequence databases. For example, the Joint Genome Institute’s Integrated Microbial Genomes (IMG) genome annotation and analysis system (Integrated Microbial Genomes 2007) uses a BLAST E-value cutoff of $10^{-5}$ as a crucial step in the annotation of predicted coding sequences (CDS, or prospective genes). Genes are assigned to clusters of orthologous groups (COGs) based on a reverse position specific BLAST (RPS-BLAST) search of NCBI’s Conserved Domain Database (Marchler-Bauer, et al. 2002) (CDD) using an E-value threshold of $10^{-2}$.

The Institute for Genomic Research (TIGR) database annotation cutoffs range from $10^{-5}$ (for putative gene annotation assignment using BLASTX against GenBank’s non-redundant amino acid database, nrna (Integrated Microbial Genomes 2007), to $10^{-10}$ (for tBLASTX mapping between potato ESTs and the A. thaliana genome (Research 2007)). TIGR’s open-source automated annotation pipeline executes a BLAST search of predicted genes, or open
reading frames (ORFs), against several protein databases, including GenBank proteins, the Protein Information Resource (PIR) and Swiss-Prot. A refined search is then performed using a modified Smith-Waterman alignment on the BLAST hits from the first step. The setting of the E-value cutoff is left to the database custodians and values as high as 0.1 have been used (Beckstrom-Sternberg, et al. 2007).

But what is a reasonable E-value cutoff? Generally, no objective criterion is given for setting E-value cutoffs, although an arbitrary threshold of $10^{-5}$ is common. Often, scientists use their experience to arrive at trusted cutoff scores and supplement these with BLAST searches against additional databases, such as EST databases. In the absence of any guidelines or previous experience, one approach is to go by the degree of human intervention in the annotation process: choose a more stringent cutoff when human intervention is absent and a more permissive threshold where human intervention is used to quality-check automated annotations.

In Chapter 2, we propose a more reliable way of setting E-value cutoffs based on an incrementalist approach using large-scale BLAST searches of the manually curated Swiss-Prot database.

3. Expressed Sequence Tags (ESTs)

Expressed sequence tags (ESTs) were the first wave of ‘genomic data’ – a convergence of genomic technology and computational power that made it possible to sample from the complete genetic composition of cells. ESTs are short (~500bp) fragments of RNA sampled
from the cell’s whole RNA population (including mRNA, pre-mRNA, siRNA, etc.).

In the sequencing process, each EST is output as a sequence of residue predictions with varying degrees of base prediction accuracy. Current sequencing technology provides a ‘quality value’ for each residue along the sequence: a measure of the likelihood that the base prediction is accurate. Overlapping clusters of reads that have very similar sequences are considered to be fragments originating from different positions along the same original RNA molecule. In an assembled collection of ESTs, a cluster of contiguous sequences is called a Contig, whilst EST strands that do not cluster with other sequences are called Singlets.

The technologies that made EST production possible, such as Sanger base-read sequencing (Sanger, et al. 1977) and nucleic acid amplification by polymerase chain reaction (PCR), and the infrastructure developed for high-throughput EST production, data management and analysis are also used to sequence whole genomes (Venter, et al. 2003). With the ‘$1000 genome’ initiative (NIH, 2006) and other advances in genomic sequencing technology, it might be thought that ESTs are no longer necessary. However, because of their origin, ESTs represent a powerful genetic analysis tool, particularly when used in concert with other genomic or clinical data (Zuo, et al. 2005), and are used in many different applications, as we shall describe in some detail below.

3.1 Condition-specific expression

Since ESTs are samplings of populations of transcripts, we can infer relative levels of expression by tissue, disease state or developmental stage by counting the numbers of
sequences making up particular gene clusters, and the numbers and types of gene clusters in any sample of ESTs. The first publicly available EST database was created in the mid-1990s to promote the exchange of genomic data, reduce duplication of effort and speed the identification of disease-related genes (Williamson 1999). This was quickly followed by the Cancer Genome Anatomy Project (CGAP), a collection over 1 million ESTs from normal, premalignant and malignant cells (Strausberg, et al. 2000).

By their nature, ESTs can only be used to discover genome translocations or rearrangements that occur within the expressed portions of genes. This limits their usefulness to questions relating to the break points within genes. However, this phenomenon has been strongly implicated in the onset or development of cancer. Chromosome aberrations are common characteristics of most human cancer cells (Mitelman et al 2004, Futreal et al 2004), which is thought to occur by the perturbation of regulatory pathways initiating tumorigenesis. The Mitelman Database of Chromosome Aberrations in Cancer contains over 45,000 examples of chromosomal aberrations associated with cancer detected by cytogenetic banding experiments (Mitelman 2006).

EST collections from individuals or organisms can be used to inexpensively detect the presence and expression of fused genes, or even incorporated into microarrays for quick assays of various cancer types. EST-based research on prostate cancer in the late 1990s led to the discovery of three genes specifically expressed in human prostate cancer (Vasmatzis, et al. 1998) and to the development of a curated relational database and suite of analysis

An extension of EST technology, serial analysis of gene expression (SAGE) was developed to provide a more cost-effective means of uniquely identifying transcripts. SAGE is based on the assumption that a 10–14 base pair stretch mRNA can uniquely identify a transcript, provided that it is derived from a unique region within that transcript. SAGE has been used to analyze developmental changes in pigment cells (Weeraratna 2003).

Spotted microarrays are another form of large-scale expression assay that could be seen as a competitor to EST technology. However, although microarrays can be used to assay the expression levels of 10,000s of genes at once, the sequences have to be first discovered (e.g., by using ESTs) and, even then, the microarray assay is ‘closed’ in that it is limited to only those genes in the array, compared to the ‘open’ form of EST gathering. Nevertheless, these three different assays can also be seen as complimentary, with ESTs usually preceding the others in the process of the genetic characterization of a phenomenon.

### 3.2 Gene identification in genomic data

EST, protein and cDNA databases were recently used to find 282 gene structures out of 1,435 predicted genes in human chromosome 12 that had not been found by automated gene finding software applications (Scherer, et al. 2006). There are at least 487 loci on
chromosome 12 that have been directly implicated in human disease.

Thus, despite the generally higher quality and coverage of genomic sequence assemblies, the relatively ‘noisy’ EST data sets are a powerful tool for identifying genes missed by current gene-finding methodologies. Depending on the depth of coverage of an EST collection, many exons may be poorly covered by ESTs or missed completely. In such cases, ESTs can be supplemented or replaced by signature sequence tags (SSTs), which are subsets of probes that can be used to detect mutants (Mironov and Pevzner 1999). However, the overall utility of ESTs and their relatively low cost has meant they remain the primary tool for investigating a novel genome.

Other uses for ESTs include transcription start site and promoter region identification, improved comparative genomics / QTL mapping (Demeure, et al. 2005) and SNP discovery and haplotyping. In an analysis of 11 genomic regions containing single nucleotide substitutions, the SNP discovery application PolyPhred had a prediction accuracy of greater than 99% with sequences generated using fluorescent dye-labeled primers and approximately 90% for sequences prepared with dye-labeled terminators (Nickerson, et al. 1997). In a later study, 5,400 novel exonic SNPs were found by aligning public EST data to the draft human genome sequence, and an additional 12,000 SNPs were found from EST cluster alignments (Hu, Modrek et al. 2002). Verification of the predicted SNPs using 20 Finnish DNA samples indicated that 82% of the genomic-aligned SNPs and 63% of the EST-only SNPs were detectably polymorphic. The authors also showed that 37% of the SNPs mapped to
known protein coding regions, leaving 6,500 novel SNP predictions. More recently, a study of 1107 SNPs on EST counts of two alleles (A and B) using a binomial test showed allelic gene expression variation in 524 SNPs (47%), in accordance with previous studies using SNP chips and primer extension (Lin, et al. 2005).

Other methods for identifying SNPs tend to build on the phred/phrap information base (Buetow, et al. 1999, Cheng, et al. 2004) with some significant algorithmic advances, such as the Bayesian approach used in PolyBayes (Marth, et al. 1999) and a more recent approach using chromatogram analysis and consensus sequences that is more accurate than PolyPhred or PolyBayes when tested on HIV datasets (Galves, et al. 2006). NovoSNP is another SNP detection method that claims a better reduction in false positives than PolyPhred and PolyBayes at high quality cutoff values (Weckx, et al. 2005). The logical extension of these approaches is to use the SNPs identified to construct haplotypes associated with tissues or disease states (Tang, et al. 2006).

3.3 Alternative splicing prediction

ESTs are considered to be completely and correctly spliced mRNAs. As such, they offer the possibility of reconstructing the splicing patterns of a gene by looking at how the different exons of a gene are combined in the sequences belonging to that gene cluster in the EST population in question. Traditional assembly methods are prone to error in alternative splicing reconstruction (Figure 1 - 2). Given that 30-60% of human genes are alternatively spliced, this could be a major source of error in the assembly process. This problem can be
avoided by using a graph representation of the assembly process, in which each transcript corresponds to a path in the graph or assembly, to arrive at a more accurate assembly and a ready-made profile of the splicing pattern of the gene cluster (Heber, et al. 2002). Based on this technique, approximately 65% of the genes from Ensembl, RefSeq, STACK, TIGR, UniGene and other major sources were found to show evidence of alternative splicing, with 5% of the genes having over 100 predicted alternate transcripts (Leipzig, et al. 2004). A database of splicing graphs for human genes using transcript information was made available, including a web interface for viewing splicing graphs and interactively assembling transcripts, and a catalog of over 1.2 million putative transcripts.

3.4 Caveat: A High Proportion of Uncharacterized ESTs

An important caveat regarding the publicly available EST databases is that there appear to be high error rates in these datasets. This is manifested by a significant portion of the available human ESTs not matching human genomic sequences. In an analysis of pairs of ESTs from dbEST which were reported to derive from a single gene, it was found that as many as 26% of the pairs do not both align with the sequence of the same gene (Wolfsberg and Landsman 1997). The authors suspected that some of these unusual ESTs result from artifacts in EST generation.

4. A short history of *F. heteroclitus*

The mummichog, *Fundulus heteroclitus*, is a euryhaline fish (i.e., it can survive both in freshwater and seawater) with an extreme temperature range between populations off the
coast of Newfoundland and along the Eastern seaboard of North America to the Gulf of Mexico. *F. heteroclitus* populations are commonly found in bays, estuaries and salt marshes and spawn in shallow water from April to August, becoming sexually mature in one year. The fish spend much of the winter months burrowed in the mud and some fish survive a second winter. *F. heteroclitus* shows striking sexual dimorphism, particularly when spawning.

### 4.1 Phylogeny

In the conventional phylogeny, *F. heteroclitus* is most similar to medaka and lies close to the two fugu species - fugu and tetraodon. These species are all more similar to each other than they are to zebrafish. A comprehensive phylogenetic analysis based on mitochondrial sequences showed that genetic data broadly supports the conventional phylogeny (Figure 1 - 3).

Knowledge of the phylogenetic relationship of *F. heteroclitus* to other fish species allows us to infer the applicability of certain comparison data and comparative approaches for particular problems. The genomes of the pufferfish species *Takifugu rubripes* and *Tetraodon nigroviridis*, the zebrafish (*Danio rerio*) and the Japanese medaka (*Oryzias latipes*) are extremely valuable resources for studying the molecular physiology of fish and other vertebrates. Phylogenetic analyses of other genes show significant similarities between *F. heteroclitus* and medaka (Kawaguchi, et al. 2005, Urushitani, et al. 2003, Nelson 1994). It is this kind of phylogenetic closeness at a single or multiple-gene level that extends the
scope of comparison to envelop basic metabolic processes in humans and other vertebrates.

4.2 *Fundulus heteroclitus* as a Genomic Resource

With its relatively short generation time (9 months to 1 year), ease of rearing and relatively inexpensive infrastructure requirements, *F. heteroclitus* is a convenient model for certain human health conditions. Lactate dehydrogenase (Ldh) is one of the genes that has been intensively studied in *F. heteroclitus*. Mutations of Ldh-A have been associated with the disease exertional myoglobinuria in humans. Ldh is one of many metabolic genes that can be studied in experiments that are relatively quick and inexpensive in this non-model organism (NMO). Furthermore, *F. heteroclitus* genomic resources provide the opportunity for comparative genomic studies aimed at identifying underlying evolutionary adaptations of closely related species that show distinct environmental responses. Further investment is required in marine genomic techniques and resources in order to integrate the results of comparative genomics to form a detailed understanding of the underlying molecular mechanisms.

As a simple example of one application of the available genomic data, candidate promoter regions identified by alignment of EST sequences against BAC bacterial artificial chromosomes of several species can be mined for motifs (Prakash, et al. 2004) or regions that show evolutionary selection or conservation. These ‘phylogenetic footprints’ (evolutionarily conserved non-coding sequences) are subject to stabilizing selection and so evolve much more slowly than adjacent non-functional DNA. The loss of phylogenetic
footprints or the acquisition of conserved non-coding sequences in particular subpopulations or species, can provide evidence for the evolutionary modification of cis-regulatory elements. BAC sequences can also be used to ‘bridge’ non-overlapping ESTs and aid in the assembly of other genomic scaffolds, thereby alleviating somewhat the complications caused by the multiple chromosomal duplications found in fish. Although chromosomal duplications can complicate EST definition and scaffold assembly somewhat, fish genomes are ideal surrogates for the investigation of gene function and allelic differences in other vertebrates, such as humans, precisely because of the frequent gene and chromosomal duplications in fish and subsequent divergence and specialization of redundant genes (Karchner, et al. 2002). In contrast to experimentally inaccessible mammalian systems which may also exhibit complicating pleiotropic effects, fish systems can be relatively transparent and easy to manipulate.

Most of the current body of research on *F. heteroclitus* has been based on wild, outbred populations. One of the advantages of a comparative genomics approach is the availability of inbred lines which dramatically reduce the noise due to genotypic variation between individuals. In addition to the environmental adaptation research discussed above, *F. heteroclitus* is used to investigate the effect of pollution and other environmental toxins and challenges because of the intimate physiological contact with all bodily fluid compartments and tissues through the gills and gastro-intestinal system.

Lastly, one major advantage of developing genomic resources for *F. heteroclitus* and other
non-model organisms is the flexibility of custom production of cDNA libraries, clone collections and microarrays. Related to this, a final argument for the development of *F. heteroclitus* genomic resources such as microarrays is the availability of eQTL (quantitative trait loci) mapping techniques that can provide useful information about a trait or physiological phenomenon of interest (Kendziorski, et al. 2006).

As outlined above, the evolutionary adaption of *F. heteroclitus* to the metabolic challenge of environmental temperature is gradually being unraveled by a multidisciplinary approach focusing on comparative genomic analyses. Building on the research data discussed above, genes evolving by natural selection in *F. heteroclitus* have been identified by measuring the expression of metabolic genes and allocating the maximum among-population variation to genetic distance and examining the remaining variation relative to a temperature (Whitehead and Crawford 2006). The variation in expression for 22% of the genes studied that regress with habitat temperature was significantly greater than could be accounted for by genetic distance alone. Evolution by natural selection was suggested as the most parsimonious explanation for among-population variation for these genes. The study showed that many metabolic genes have patterns of variation that cannot be explained by neutral evolution, i.e., too much or too little variation, which may reflect important physiological functions.

Besides Ldh-B and other *F. heteroclitus* metabolic genes that have orthologues in humans and other vertebrates, *F. heteroclitus* also expresses the cystic fibrosis transmembrane conductance regulator (CFTR) in its gills and intestinal epithelia (Marshall and Singer 2002).
A phylogenetic comparison of the CFTR gene in several teleost species showed all the major elements of the human gene are present and the ability of teleosts to rapidly regulate CFTR expression and activation during salinity adaptation makes the fish an attractive model for studying the expression and cellular transport of CFTR.

A recent paper indicated that the heat shock response in *F. heteroclitus* involves a complex dynamic of multiple isoforms (Fangue, et al. 2006). The paper examined variation in whole-animal thermal tolerance and differences in the sequence or regulation of the heat shock protein genes in extreme northern and southern *F. heteroclitus* populations. The authors found no fixed differences in amino acid sequence between populations but significant differences between populations in the expression of many of the genes. This complements an earlier finding that the Sp transcription factor domains have evolved independently (Kolell and Crawford 2002) according to the phylogenetic characterization of Sp in mammals and fSp3, the Sp protein expressed in the liver or heart tissue of *F. heteroclitus*. The results suggest that Sp2 should not be considered a member of the Sp family and that only two domains (zinc fingers and B domain) share similarity outside the Sp family. The zinc fingers are homologous to other GC-binding domain and the B domain is homologous to protein-protein interacting domains in the CCAAT-binding/NF-Y transcription factor families, suggesting these different domains have different evolutionary histories.

The ongoing development of microarray resources for *F. heteroclitus*, which can be used in other teleost species, is an important part of the comparative genomics paradigm in which
interesting physiological traits and responses that are experimentally intractable - or absent in model species are better illustrated in non-model organisms. Microarray technology can be successfully applied to non-model species to ameliorate the current poor understanding of the system-wide environmental stress response in complex, multi-tissue animals (Gracey, et al. 2004). This can either be in the form of hypothesis-driven or discovery-based multi-locus scans, comparing different populations to identify variations in genomic regions that result in environmental adaptation (Podrabsky and Somero 2004).

Oligonucleotide microarray analysis also can successfully identify many potential direct downstream genes and decipher downstream signaling pathways (Chen, et al. 2005). The temporal pattern of gene expression in response to environmental change can now be investigated using cDNA microarrays. Traditionally, gene expression analysis has focused on fold-change analysis, because conventional techniques for time series analysis are ill-suited to short series of time-ordered data. However, techniques are available to infer statistically meaningful information from a small set of four or more time-ordered gene expression measurements (De Hoon, et al. 2002). Finally, another area where comparative analysis using *F. heteroclitus* can provide biological insight is the exploration, development and refinement of different methods of using phylogenetic information in comparative studies (Garland, et al. 2005).

In summation, *F. heteroclitus* EST resources can provide an important resource in marine and non-marine genomic research, including studies related to human health, population
genetics and comparative genomics. As we shall discuss in the following chapters, better quality control and novel uses of sequence similarity searches and protein motifs can contribute to improving the annotation of EST databases and enhance the benefits of their use with other genomic resources.

5 Machine learning using protein motif inputs

The third chapter of this dissertation involves the use of machine learning techniques to predict protein function. One challenge of using computational techniques such as machine learning to determine underlying patterns in biological data is retaining the biological sense of the original data inside the computational model such that the result is readily understood in biological terms. Machine learning techniques such as neural networks and support vector machines are commonly used for determining patterns in numerical data. However, the ability to use non-numeric data is crucial in our choice of machine learning algorithm for the classification of protein motifs; although many techniques exist for transforming non-numeric data to numeric values, they cannot be meaningfully applied to signatures of protein motifs in the form of protein motif names. Decision Tree and Naïve Bayes machine learning approaches were chosen for this study because they can use non-numeric inputs such as protein motif names and the resulting models can be easily understood by the untrained eye. The relatively straightforward interpretation of Decision Trees in particular makes them particularly suitable for the pedagogic function of improving our understanding of the biological dynamic between sequence and function.
The advantages and disadvantages of Decision Trees are summarized in Table 1 - 1. One chief advantage of the Naïve Bayes method is that prediction results are easy to interpret: Naïve Bayes produces a probability score ranging from 0 to 1 for each category and the category with the highest probability score is considered the ‘correct’ choice. Similarly, Decision Trees are bifurcating trees of rules with probabilities at each branch point that provide a decision as to the identity of the sample being tested and also a statistical indication of the likelihood that the prediction is correct.

5.1 An introduction to Decision Trees

For a brief introduction into how Decision Trees work, take a Decision Tree on whether or not to play golf: It can be easily understood as an upside-down tree of choices where certain attributes are forks in the branches Figure 1 - 4. Here, the attributes (or forks) available are: outlook, temperature, humidity and wind. The branches splitting off from the forks are the different values of attributes, e.g.: outlook has three values: sunny, overcast and rain. Depending on the conditions on any particular day, the Decision Tree will be traversed down its branches to the final leaf - the decision on whether or not to play that day. Similar to the golf Decision Tree, the decision on whether or not a protein belongs in a particular functional group can also be represented as a Decision Tree.

Note that not all attributes are necessarily considered important in the construction of a Decision Tree and some may be left out of the tree. For example, in the golf Decision Tree, the temperature attribute is not used because it does not provide sufficient information
regarding the previous outcomes to make it a useful deciding factor compared to the other attributes. The relative importance of attributes is commonly allocated based on information gain or the Gini impurity. N-fold cross-validation tests are commonly used to validate Decision Trees models and provide a measure of their reliability.

5.2 An introduction to Naïve Bayes

The Naïve Bayes machine learning algorithm is a well-studied probabilistic algorithm often used in automatic text categorization. It is much faster than other popular machine learning algorithms such as k-Nearest Neighbors, Support Vector Machines or Decision Trees, and it also performs very well compared to these algorithms, despite its ‘naïve’ assumption of independence between attributes. In fact, Naïve Bayes has been shown to perform as well as or better than more powerful classifiers for commonly seen set sizes and numbers of attributes, even under conditions where attributes are clearly dependent, such as linguistic analyses (Domingos and Pazzani 1997). Naïve Bayes classifiers have been used in the prediction of protein-DNA interactions with 44% specificity and 41% sensitivity in identifying interface residues using the identities of a window of 9 residues and the sequence entropy of the target residue as input (Yan, et al. 2006), where the Naïve Bayes classifier performed better than searching for PROSITE DNA-binding motifs.

In mathematical notation, the probability of event A occurring (or being true) given that B has occurred (or is true) is denoted as P( A | B ), a conditional probability, and the probability of A occurring is denoted as P( A ).
Bayes' Theorem is a way of inverting a conditional probability. It states:

\[ P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \]

When considering, for example, allocating documents to one of several categories based on the frequencies of words in the document, we are only interested in the maximum size of the numerator because this represents the greatest probability that a given document belongs to a particular category:

\[
Best\ category = \text{ArgMax}_{\text{category in categories}} \frac{P(\text{words} \mid \text{category}) \cdot P(\text{category})}{P(\text{words})}
\]

We can drop the \( P(\text{words}) \) term since it doesn't change over the range of categories, to arrive at:

\[
Best\ category = \text{ArgMax}_{\text{category in categories}} P(\text{words} \mid \text{category}). P(\text{category})
\]

Finally, the naïve assumption occurs if we denote the words in the document as \( w_1, w_2, \ldots, w_n \), and say that the equation above is equivalent to:

\[
Best\ category = \text{ArgMax}_{\text{category in categories}} P(w_1 \mid \text{category}) \cdot P(w_2 \mid \text{category}) \cdots P(w_n \mid \text{category}) \cdot P(\text{category})
\]

This last step naively assumes that the attributes, in this case, words in a document, are independent. This is clearly not the case as a word might tend to be commonly used with other words, so the presence or absence of these other words would affect the likelihood of finding this word. For this study, the naïve assumption is that the occurrence of a protein
motif in a particular protein does not affect the chance of any other motif also being found in the protein.
References


55 Pevzner Group, B.L., University of California, San Diego, Splicing graphs, Available at [http://www.cse.ucsd.edu/groups/bioinformatics/ESTs/index.html](http://www.cse.ucsd.edu/groups/bioinformatics/ESTs/index.html).

Table 1 - 1 The advantages and disadvantages of Decision Trees and Naïve Bayes.

Compared to other methods, such as neural networks and support vector machines, Decision Trees and naive bayes provide similar performance and produce output parameter values that are easy to extract and understand.

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Decision Trees</th>
<th>Naïve Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intuitive</strong> - Easy to understand and interpret after a brief explanation</td>
<td><strong>Simple</strong> - A basic probability calculation.</td>
<td></td>
</tr>
<tr>
<td><strong>Real values</strong> - Can use both numeric and categorical data, including 'real values' (such as protein motif names)</td>
<td><strong>Real values</strong> - Can use both numeric and categorical data, including 'real values' (such as protein motif names)</td>
<td></td>
</tr>
<tr>
<td><strong>‘Glass box’ model</strong> - Provides clear, boolean inferences using the input attributes themselves, unlike the ‘black box’ model of other machine learning method.</td>
<td><strong>Good performance</strong> - Even when the assumption is independence is not met, the performance of Naïve Bayes models is comparable to more complex methods.</td>
<td></td>
</tr>
<tr>
<td><strong>Cheap and Robust</strong> - Relatively little data manipulation is required to build a decision tree model. The algorithm scales up well with large data sets and runs fast on generic hardware.</td>
<td><strong>Cheap and Robust</strong> - Scales up well with large data sets and runs fast on generic hardware.</td>
<td></td>
</tr>
<tr>
<td><strong>Simple</strong> - Relatively little data preparation – no data normalisation, assumed priors or removal of blank values</td>
<td><strong>Simple</strong> - Relatively little data preparation – no data normalisation or removal of blank values</td>
<td></td>
</tr>
</tbody>
</table>

| Disadvantages                                                                 |                                                                                         |                                                                                                 |
| **Complex rules** - Large and complex trees may produce rule sets that are difficult to follow | **Decision balance bias** - Requirement for balanced data sets (equal positive and negative examples) |                                                                                                 |
| **Suboptimal trees** - Overfitting can produce suboptimal trees that are overly complex, i.e., more suited to the training set than the generalised example | **No optimization** - No parameters to adjust to maximize performance – it’s just a simple reflection of the probabilities |                                                                                                 |
| **Supervised learning only** - Decision trees require labelled data (e.g., ‘true’ and ‘false’) and as such a supervised learning approach. | **Missing values hack** - Zero probabilities or missing values require mock priors (e.g., Dirichlet Prior) |                                                                                                 |
Figure 1 - Orthologues and paralogues are two types of homologous sequences.

Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function. Paralogy describes homologous genes within a single species that diverged by gene duplication. (NCBI 2007)
Figure 1 - 2 Traditional assembly methods often give incorrect results because they fail to consider alternative splicing. (Pevzner Group)
Figure 1 - Distribution of genome sequencing, major EST projects (>1000 sequences), large-insert libraries and radiation hybrid panels in relation to a conventional fish phylogeny.

The colored boxes, ovals and circles to the right indicate genome-sequencing projects (in purple), ESTs (in green, where size indicates the number of ESTs in publicly available collections), large-insert libraries (mainly BACs, in red) and radiation hybrid panels (in yellow) (Cossins and Crawford 2005).
The golf Decision Tree attempts to decide whether or not to play based on the current weather conditions combined with previous decisions to play or not play on different days. Each row in the table represents a day and each column a weather condition, with the last column showing the decision on that day (Quinlan 1987).
Chapter 2: Incremental mismatch probability filter improves BLAST annotation accuracy
Abstract

New collections of sequences are annotated based on DNA sequence similarity. Potential problems are that the resulting annotations are uninformative, non-standardized, incorrect or lacking in essential features that inform scientists. Depending on the database, between 12% to 85% of annotations are of poor quality. When searching other UniGene mammal databases with well-annotated Human UniGenes, between 3% to 7% of these genes are missing. Yet if there was a significant BLAST hit, these Human UniGenes match four or more sequences at E-values less than $10^{-100}$. More problematic was that the annotations were often dissimilar: for a hit with an E-value of $10^{-5}$, 69% of sequences had annotations that differed from the human sequences. This proportion declined to 4% at E-values less than $10^{-201}$. Using a Fuzzy-matching algorithm, one can determine whether annotations of the best hits are similar. If best-hits in different databases are the same, they correctly identified the original annotation 99% of the time. A second computational approach is to calculate an incremental mismatch probability that indicates the probability that the functional annotations for the query and target sequences are the same given a hit with a particular E-value (i.e., a conditional probability). Using very low E-values mitigates many of the problems with annotation, but increases the possibility of missing orthologous sequences. Alternatively, by using FuzzyMatch to compare the annotation within and among databases, one can provide a statistic for the likelihood of a correct annotation that significantly reduces the rate of incorrect annotations.
Introduction

Annotating genomic coding regions, cDNAs or ESTs is problematic (Crawford 2007, Fields and Johnston 2002, Tamames and Valencia 2006, Wain, et al. 2002). The first problem is that there are few widely-accepted rules for naming genes. Thus, most genes have several names (e.g., 53% of Swiss-Prot genes have 2 or more synonyms) and many names have different genes associated with them. For example, RAD refers to both “Ras-related associated with diabetes” and DNA damage repair gene. In mammals GRIM is an enzyme subunit in oxidative phosphorylation that is also associated with cell death; yet, in Drosophila GRIM is an unrelated cell death associated protein without enzyme function. To address these problems, the Human Genome Organization (HUGO) initiated a heroic effort to provide a single official symbol for genes (Wain, et al. 2002). This single annotation would have provided a common vocabulary for researchers in which different functions could be appended. Unfortunately, too many HUGO terms have been arbitrarily replaced one or more times by non-standard terms (Tamames and Valencia 2006). Without a standard vocabulary, integrating information from different publications into meaningful functional genomics becomes much more difficult (Crawford 2007).

The second problem relates to annotating genes based on sequence similarity. Most annotation schemes use the best BLAST hit, or hit with the lowest E-value to assign a name to the query sequence. The BLAST E-value is defined as the expected number of non-homologous sequences with a certain score in a database of a particular size. E-values are not equivalent to the likelihood of a chance match between functionally distinct sequences.
BLAST E-values as low as $10^{-50}$ against targets with different enzyme functions can occur (Rost 2002). In addition to the problems of interpreting E-values, the difficulty of evaluating BLAST hits is compounded by a lack of a unified vocabulary for describing genes. The assumption is that the best hit is homologous to the query sequence. This homology (defined as descended from a common ancestral gene (Fitch 2000)) between genes is assumed to be indicative of functional similarity. Yet, many annotations use BLAST searches with upper-threshold E-values of $10^{-1}$ to $10^{-10}$ (Pascal and Tjian 1991, Adams, et al. 1991, Clark, et al. 2003, Lee, et al. 2007). Additionally, even in the most highly investigated genomes there is a significant frequency of genes annotated with “similar to” or “predicted”: for human, mouse, rat or dog UniGenes at NCBI these frequencies are 8%, 9%, 24% and 53% (respectively). Annotation based on these approaches without some form of verification could create a cascade of errors.

This study quantifies the problems with annotation and explores the utility of the “best match” approach to naming genes. The data presented here demonstrate that even for vanishing small E-values (i.e., E-value = $10^{-100}$) there are many misannotations (annotations that are dissimilar to the original UniGene annotation). We suggest a solution for these problems using fuzzy logic and similarities between genes within databases and suggest that genes be identified with reference to two or more common databases, for example Human or Mouse UniGene or Swiss-Prot.
Methods

2.1 UniGene and Swiss-Prot BLAST Comparisons

For BLAST comparison purposes, 1,000 well-annotated Human UniGene sequences (hereafter referred to as the ‘human 1,000’ sequences) were selected from NCBI UniGene Homo sapiens release #196 and WU-BLAST BLASTN alignments were carried out against three mammalian UniGene databases: *Canis lupus familiaris* (dog), *Mus musculus* (mouse) and *Rattus norvegicus*. BLASTX alignments were also carried out against Swiss-Prot. In all cases, we used the full-length mRNA for our similarity searches. We used two criteria to select ‘well-annotated’ sequences. First, we excluded sequences with annotation that appeared to be based on sequence similarity alone or where the annotations were uninformative, i.e., genes with annotations containing the words ‘similar to’, ‘predicted’, hypothetical, ‘unknown’, ‘transcribed locus’, ‘clone’, ‘containing’, ‘like’, ‘repeat domain’, ‘domains’, ‘homolog’, ‘unknown’, ‘amplified sequence’, ‘non-coding RNA’ or ‘associated’. Second, we manually inspected annotations for specific and emphatic reference to gene function (i.e., annotations with just clone number or chromosomal location, or ambiguous terms were excluded).

2.2 BLAST scores, Histograms and E-value

Only BLAST hits less than E-value $10^{-5}$ are used. The maximum number of hits per query was 250, and low complexity regions were filtered using seg. For histograms, each bar represents the number greater than the preceding bar value and equal to the value listed for
that bar. For example if the X-axis is E-Value with -20 and -30, then a bar value of -30 indicates all hits with E-values between $10^{-21}$ and $10^{30}$.

### 2.3 Fuzzy Matching of BLAST Annotations

To determine similar and dissimilar annotations, a simple fuzzy match (Zadeh 1965) heuristic was developed. The fuzzy match heuristic involved breaking down each pair of input sentences into their component words, then comparing all combinations of word pairs containing one word from each set. To increase the effectiveness, some general terms are excluded (e.g., chain, domain, fragment, family). Each word pair match is given a score from 0 to 1, which was calculated as follows: the number of 3-letter fragments of the shorter word that are found in the longer word, divided by the length of the shorter word less the window width. The highest word pair scores for each word in the sentence with least words were added together and divided by the number of words in the shorter sentence to arrive at a match score. If the shorter sentence contained more than one word, the above procedure was repeated with one word alternately removed from the shorter sentence to arrive at additional match scores. The highest of the match scores was used as the fuzzy match score. Due to the nature of the fuzzy match algorithm, shorter, more general descriptions were more likely to be predicted as similar (e.g., ‘protein kinase’ compared to ‘serine/threonine protein kinase’). This was not necessarily a drawback with regard to evaluating the performance of the fuzzy match algorithm as more general annotations could be considered incomplete but not incorrect descriptions of the protein function.
Verification of fuzzy matching was accomplished by manually determining annotation similarity and comparing this to fuzzy match predictions. BLAST hit annotations were considered similar to the query sequence annotation if both annotations described essentially the same protein function without the need to refer to external sources such as expert knowledge of the proteins to impute associations between the annotations. Different substrates for enzymes with the same basic function were considered to be dissimilar, whereas different subunits of the same protein were considered to be similar. The Type I and Type II errors using different similarity scores were calculated based on the manual checks. The false positive rate is the proportion of negative instances that were erroneously reported as being positive (FP / (FP + TN), where FP = false positives and TN = true negatives). The sensitivity of a test is equal to 1 - the false positive rate. The false negative rate is the proportion of positive instances that were erroneously reported as negative (FN / (FN + TP), where FN = false negatives and TP = true positives). The specificity of a test is equal to 1 - the false negative rate.

### 2.4 Incremental mismatch probability to validate BLAST hits

To provide a score for the likelihood that a BLAST match gives the correct annotation, we calculated an incremental mismatch probability for the best hit of each query sequence. The best hit for a query is BLAST against its own database and the similar and dissimilar annotations determined for the best hit query and its targets. The incremental mismatch probability is the ratio of dissimilar annotations to the total number of annotations at or below a particular E-value. If the E-value for the query sequence’s best hit is intermediate
(relative to the lowest and highest E-value for self-BLAST) then the mismatch probability calculation used all lower E-values (those with greater significance), and the next higher E-value multiplied by a prorated value between 0 and 1 which is calculated as the relative distance between the next lowest and the next highest value. Distance is measured as the natural log of the E-value, thus \((\ln Q - \ln L)/(\ln H - \ln L)\), where, relative to the best hit, \(Q\) is the query E-value, \(L\) is the next lowest E-value and \(H\) is the next highest E-value in the best hit’s self-BLAST distribution. This calculation works for all query E-values except those that exceed all BLAST E-values (i.e., the query E-value is less significant than all self-BLAST) or if query E-values is the most significant. If the query E-value was the most significant, then the annotation for the best hit is assumed to be correct with a mismatch probability equal to the E-value. If the query E-value is higher (less significant) than all the E-values of the self-BLAST distribution, the mismatch probability is defined as the ratio of total dissimilar annotations/total annotations. The E-value of the original query-target best hit is then compared to the distribution of E-values for the target’s similar and dissimilar annotations. The incremental mismatch probability for the query-best hit match is defined by where the E-value of the query-best hit falls within the distribution of similar and dissimilar hits.

Results

In a perfect world, all transcribed RNAs in a genome would be identified and genes would be annotated using standardized terminology. If all messages were similarly annotated then the annotation from the lowest E-value of a BLAST search would more likely be indicative of both functional and evolutionary similarity. To measure the completeness of the UniGene
databases, we randomly selected 1,000 well-annotated Human UniGenes (human-1,000). Genes with “similar to”, “predicted”, “hypothetical”, or “unknown” and other similar terms (see methods) were excluded. To ensure we were searching using “good genes”, we also manually examined the annotations so that each annotation included a specific and emphatic reference to gene function or role (i.e., we excluded genes with just gene number or chromosomal location or ambiguous terms).

3.1 Estimated proportion of missing orthologues

We carried out BLAST searches of the human-1,000 against three other mammalian UniGene databases (dog, mouse, rat) and Swiss-Prot (Table 2 - 1). The percent of well-annotated Human UniGenes that had matches in any one of the three mammalian UniGenes with E-values of $10^{-5}$ or less ranged from 81% to 93%. Pooling the three mammalian UniGenes increases the percent matches to 97%. Ninety-three percent of the human-1,000 sequences were found in Swiss-Prot. Thus among mammals with genomes and active research communities using these genomic resources, approximately 3% to 7% of well-annotated human genes are missing. Although these “missing” genes could be human-specific and not present in other mammals’ genome, it is more parsimonious to suggest that these well-annotated Human UniGenes are not unique, but instead have not been identified in these other mammals. This supposition is supported by the finding that 98% of the human-1,000 genes have at least one hit when combining all four databases (Table 2 - 1).
3.2 Estimation of annotation quality

Interdependence of annotations, where annotation is primarily based on sequence similarity, creates problems if the original annotation is ill-conceived. To determine how frequently annotations are dependent on other annotations, we examined the annotations from the BLAST of the human-1,000 for the term “similar”, “predicted”, or “hypothetical”. Additionally, we searched for annotations that provided little functional information but instead had only clone numbers, sequence read ID, or an obscure phenotype (e.g., “Groucho”). For genes similar to the human-1,000, 7% to 72% of hits were dependent (Table 2 - 2) in the four databases. Poor quality annotations (Table 2 - 2), including dependent annotations and annotations lacking a description of gene function, were particularly abundant in the rat and dog (70% to 85%, respectively). Notice that the manually curated Swiss-Prot database has the lowest percent of “Poor quality” annotations even though it is the largest database with the most number of hits.

The other parameter of interest is the number of significant hits each query matched. For all the human-1,000 sequences that had a hit in Swiss-Prot or the three mammalian UniGene databases, we interrogated the data to discern the number of hits per query with E-values less than $10^{-5}$, $10^{-50}$ or $10^{-100}$ (Figure 2 - 1). In general, the less stringent the E-value the more hits per query. Thus, at an E-value of $10^{-5}$, 84% of queries had more than 10 hits, and 36% of queries had 100 or more BLAST hits against the combined Swiss-Prot and mammalian UniGene databases (Figure 2 - 1C). Even with an extreme E-value cutoff of $10^{-100}$, the majority of queries had 4 or more hits when BLAST against either Swiss-Prot (55%; Figure 2 - 1A) or the combined databases (65% Figure 2 - 1C). For the three mammalian databases at
E-values of $10^{-50}$ or $10^{-100}$, 26% and 13%, respectively, of the human 1,000 queries had 4 or more hits.

### 3.3 Estimating the similarity of E-values

There are two relevant questions concerning the multiple matches per query: 1) how similar are the E-values for the different matches and 2) how similar are the annotations. For the three mammalian databases, we examined the second, third and fourth best hits for each query sequence and compared them to the best hit as the ratio of natural logs of E-values (Table 2 - 3). The second hit had an exponent that was >95% of that of the best hit for two hundred and forty-four (26%) of the human queries. More than 90% of the query sequences have a second hit where the exponent of the E-values is 50% of that of the best hit (Table 2 - 3). Thus, for example, when the best hit has an E-value of $10^{-100}$, in 26% of cases the second hit will have an E-value of $< 10^{-90}$, and in more than 90% of cases the second hit has an E-value of $< 10^{-50}$. Similarly, 33% of the third hits are within 20% of the best hits. Most researchers would gladly accept an annotation based on an E-value found for the second or third hit. If one combines the data from figures one and two and table 3, many genes will have multiple hits with E-values less than $10^{-100}$, and the difference in E-value between the best and next best hits is relatively trivial (e.g., $10^{-100}$ versus $10^{-90}$).

Four or more hits with similar E-values would be fine if all homologous genes were annotated with similar names and were defined for similar function. Yet, based on a human versus other mammals comparison, a significant minority (7-19%) of well-annotated human genes will be missing from one or more of the mammalian databases. With many matches
per human gene and many orthologous genes missing altogether, the best hit may come from a less closely related species and may be less likely to have the same function as the query gene. Furthermore, if annotation depends on more evolutionary distant organisms, there is a potential for functional differences or differences in the form of the annotation. For example, although Drosophila has 32 nuclear subunits for the first protein-complex of the oxidative-phosphorylation pathway (NADH dehydrogenase) only four are annotated such that they suggest this function (FlyBase). The other names for NADH dehydrogenase subunits in Flybase include neo18/lethal or lack any functional information (e.g., CG3446). These annotations can be misleading in the context of other databases.

3.4 Fuzzy Match Algorithm matches similar annotations

How similar are the annotations for sequences with a similarity that produces extreme E-values (less than $10^{-100}$)? To automate the comparison of annotations we developed a ‘FuzzyMatch’ algorithm which evaluates the word similarity of biological annotations (Figure 2 - 2). FuzzyMatch provides a score from 0 to 1.0 that describes the similarity of annotations (see methods). To evaluate the effectiveness of FuzzyMatch, we manually compared the annotation among Swiss-Prot to the human 1,000 annotations (total of 870 Swiss-Prot hits with functional annotations, Table 2 - 4). At a FuzzyMatch score cutoff of 0.5, the false positive rate (Type I = FP / (FP + TN, where FP = false positives and TN = true negatives) and false negative (Type II = FN / (FN + TP, where FN = false negatives and TP = true positives) are 3.6 and 5.9% respectively. Assuming these error rates, we use FuzzyMatch scores of 0.5 or greater to accept annotations as similar enough to be equivalent
functional descriptions. FuzzyMatch scores below 0.5 were labeled dissimilar or mismatched.

Using FuzzyMatch, we compared the annotations between human-1,000 and Swiss-Prot. For Swiss-Prot hits with an E-value of $10^{-5}$, 31% (3,088) have similar annotations to the human UniGene query genes and thus 69% (6,917) were dissimilar (Figure 2 - 3A). Not surprisingly, the relative frequency of similar annotations increases with more significant (lower) E-values. Thus at E-values of $10^{-50}$ or less, more than 82% of hits have annotations that are similar to the original Human UniGene annotation. To investigate the relationship between the proportion of similarly annotated hits and the total number of hits for each query sequence, we plotted the number of similar and dissimilar hits with an E-value of $< 10^{-100}$ versus the number of hits per query (Figure 2 - 3B). Eighty-six of the human queries had only a single hit and, of these, 12.8% had annotations that were dissimilar to the original human annotation. For query genes with 2 to 200 hits, between 7.1% and 21.3% of annotations were dissimilar (Figure 2 - 3B). These data suggest that the probability of accepting a misannotation with a very low E-value is independent of the total number of hits for the individual query gene.

To provide a broader comparison of annotations, all Swiss-Prot genes were BLAST against Swiss-Prot, thus there was one perfect match for each query. There were approximately 18 billion hits for the 188,477 queries (Figure 2 - 4). For the 3.46 billion BLAST hits with E-values between $10^{-5}$ to $10^{-10}$, 37.9% (1.31 billion) of the annotations were dissimilar and 62.1% (2.15 billion) had similar annotations. For the 1.6 billion hits with E-values between $10^{-50}$ to $10^{-100}$, 5.4% or less had dissimilar annotations. The frequency of mismatches
decreases with greater E-value, but never disappears even at E-values as low as $10^{-300}$. These
data are similar to Human 1,000 versus Swiss-Prot at more significant E-values. The greater
proportion of matches at less significant E-values, most likely reflects the more uniform
annotation scheme employed by Swiss-Prot in which less characterized genes are given
similar annotations to more finely characterized genes, e.g., ‘protein kinase’ versus ‘serine-
threonine protein kinase’.

To evaluate the utility of using multiple databases to increase the accuracy of taking the
lowest E-value hit from among BLAST hits, we manually compared the annotations for the
best hits for the human 1,000 UniGenes BLAST against Swiss-Prot and Mouse UniGene
(Table 2 - 5). For the 1,000 human queries, 886 had a match in both Swiss-Prot and Mouse
UniGene. There were three groups of annotations (gray rows, Table 2 - 5): 1) For 374 (42%)
human UniGenes with matches against both databases, the best hits in both Swiss-Prot and
Mouse UniGene had similar annotations, 2) For 133 query genes (15%), the two best hits in
these databases had dissimilar annotations, and 3) In the case of 379 query genes (43 %), one
or both of the best targets lacked functional information (i.e., were only annotated with clone
number, chromosomal location, etc.). Most of the non-functional annotations (99.5% of 379)
are from the Mouse UniGene. The most interesting observation is that if both annotations
were similar they were also similar to the human query annotation 99% of the time.

Compare this to where the annotations for the two targets were dissimilar: only 68% of the
time was the annotation of the most significant (best) E-value target match similar to that of
the human query and, 23% of the time, neither were similar to the query annotation. These
data suggest that the accuracy of annotations can be increased by comparing the results of sequence similarity searches among multiple databases. Most sequences have multiple significant hits and many of these are annotated differently. One implication of this is that genomic projects that rely on sequence similarity–based computational annotation rather than experimental evidence will inevitably have incorrect annotations. The frequency of incorrect annotations is related to the relative similarity or E-value for a BLAST search. At low, but commonly used E-values of $10^{-5}$, or $10^{-10}$, approximately 70% to 60% of hits can have incorrect annotations (Figure 2 - 3). Although the frequency of incorrect annotations decreases with lower E-values, it almost never completely disappears.

3.5 The utility of the incremental mismatch probability statistic

We have arrived at a very similar conclusion to that of previous research regarding the utility of E-values in determining annotation matches: even very low E-values cannot be relied upon as global cutoffs for the purposes of annotation. The difficulty of connecting E-values to match predictions is due to the lack of information regarding the degree of similarity of the target to other sequences with the same or different functions and different BLAST E-values. Furthermore, this problem is multiplied when trying to compare different targets from the same or different databases with similar E-values but different annotations. To provide some of the missing information regarding the E-value distribution of functionally matching and non-matching sequences relative to any particular target, we defined an “incremental mismatch probability” statistic for identifying incorrect annotations. This statistic is a
estimates the probability that a BLAST hit target sequence does not share the same function as the query sequence. Primarily, the incremental mismatch probability measures the ‘fitness’ of a BLAST match for the purposes of evaluating whether or not to transfer functional annotation from the target sequence to the unknown query sequence. The incremental mismatch probability employs the actual distribution of match and mismatch BLAST hits for a given target sequence from among a large database of sequences. It is a conditional probability because it estimates the probability that a BLAST query sequence does not share the same annotation as a target sequence given the particular E-value of their alignment. Calculating the incremental mismatch probability involves fitting a simple distribution function to the mismatch/match ratio across a range of E-values (defined as from $10^{-5}$ to the lowest E-value in the distribution of BLAST hits).

The incremental mismatch probability is mathematically defined as follows. For each unknown gene and its best BLAST hit in a particular database we are interested in testing the hypothesis that the target has a different function from the unknown gene:

$$H_0: x = 0 \text{ versus } H_1: x = 1$$

where $x = 0$ i.e., the sequences have dissimilar functions, as opposed to $x = 1$ the sequences share the same function.

Given a BLAST match between an unknown sequence $A$ and an annotated sequence $B$ with a particular E-value $E$, the probability that $A$ shares the same function as $B$ is represented by the following conditional probability mass function:

$$f(x|e) = P(X=x|E=e,I)$$
where $X = \{0,1\}$, and as above

$x = 0$ i.e., the sequences have dissimilar functions, as opposed to

$x = 1$ the sequences share the same function, and

$I = \text{the distribution of E-values of all BLAST hits for B in a large and unbiased}
\text{sequence database (with E-values ranging from } 10^{-5} \text{ to the lowest E-value among the}
\text{hits). This distribution is an empirically calculated discrete probability distribution}
\text{(Figure 2 - 5; see methods) whose cumulative distribution function increases in jumps}
\text{each time a BLAST hit is encountered along the range of E-values.}

An example of a mismatch probability distribution would be the two values $f(0|10^{-100}) = 0.05$
and $f(1|10^{-100}) = 0.95$ denoting the probability of a mismatch and a match, respectively, given
an E-value of $10^{-100}$. An example of how the incremental mismatch probability is calculated
for single human UniGene S15967514 (Cytochrome c oxidase subunit 8A) BLAST against
human UniGenes is shown in Figure 2 - 5. As the E-value increases (is less significant) there
are more hits. At a cumulative number of 20 hits (orange dotted line), a mismatch hit
(Glutaminyl-tRNA synthetase, red diamond) occurs with an E-value of $2.0 \times 10^{-25}$, raising the
incremental mismatch probability to 0.05 (blue line Figure 2 - 5). The incremental mismatch
probability rises again beyond 0.05 due to a mismatch hit with E-value $2.0 \times 10^{-7}$. Assuming
a mismatch probability threshold of 0.05 or less, any hit against human UniGene S15967514
with an E-value or less than $2.0 \times 10^{-7}$ would be annotated as 'Cytochrome c oxidase subunit
8A'.
The incremental mismatch probability is distinct from the BLAST E-value and WU-BLAST p-value in both its underlying assumptions and the property being estimated. To explore these differences we must first understand the properties of the E-value and p-value statistics. The primary motivation behind the development of the E-value was the fact that BLAST match scores do not reflect database size. The Karlin-Altschul equation (Karlin and Altschul 1990) states that the number of alignments expected by chance (E) during a sequence database search is a function of the size of the search space (m. n), the normalized score (λS), and a minor constant (k):

$$E = kmne^{-\lambda S}$$

The E-value of a BLAST alignment of two sequences with score S is thus the expected number of alignments to be found with score S or higher in two random sequences with the same lengths and base or peptide compositions. Because E increases linearly with the number of database sequences, a BLAST hit in a database of 1,000 sequences will be 1,000 times more significant than a BLAST hit in a database of 1,000,000 sequences.

The p-value of an E-value reported by the WU-BLAST algorithm is found using the Poisson formula:

$$p(E) = 1 - e^{-E}$$

For two sequences, A and B, with a BLAST alignment score S, the p-value is the likelihood that two random sequences with the same lengths and base or amino acid compositions will align with a score greater than or equal to S (Misener and Krawetz 1999). For E-values less than $10^{-5}$, the p-value is approximately equivalent to the E-value.
Although p-values are commonly used in statistics to describe statistical significance, the E-value is the primary measurement used to assess BLAST hits. The principal assumptions behind both of these statistics are:

1. The distribution of amino acids or nucleotides in the target database and query sequences is random
2. The probability of a base or amino acid (letter) appearing at any given point in a sequence is independent of the letters at other positions in the sequence
3. Matching regions along the query sequence are independent of each other, and similarly for the target sequence
4. The database sequences are evolutionarily unrelated and thus independent of one another.

Thus the object of the E-value and p-value statistics is to estimate the propensity of random sequences to align by chance with particular alignment scores, whereas the object of the incremental mismatch probability is to estimate whether or not the sequences of such an alignment share the same function. In comparison with these other statistics, the incremental mismatch probability statistic was found to be somewhat successful in evaluating BLAST hits between sequences with matching annotation and hits between sequences with functionally different annotation. This may be due to the more relaxed assumptions implicit in the incremental mismatch probability statistic. One of the assumptions of the incremental mismatch probability statistic is that the BLAST alignment search space is complete and equivalent to, or at least an unbiased sample of, the set of all existing proteins. The larger the target database, the more reasonable this assumption becomes. Our findings are based on
Swiss-Prot, the largest manually annotated protein database. Aside from this criterion, the incremental mismatch probability statistic assumes that database annotations are complete and correct but is far less stringent than the E-value with regard to the nature of the sequence databases. In particular, the incremental mismatch probability does not require the four major assumptions for E-values and p-values, namely randomness of letters and independence of letters in different positions, different match regions and different database sequences.

We calculated an incremental mismatch probability for the best Swiss-Prot hits for 852 human UniGenes where the Swiss-Prot hits had functional annotations and manually compared the original human UniGene annotation to the best hit Swiss-Prot annotation to determine the frequency of type I and type II errors for the incremental mismatch probability method (Table 2 - 6). For the 852 human queries, 794 of the Swiss-Prot best hit annotations were similar to the human annotation based on manual inspection and 58 had different functions (i.e., they were false positives). Perhaps due to its reliance on an imperfect fuzzy match algorithm or due to errors in database annotations, the incremental mismatch probability has a relatively high Type I error rate of approximately 50% (the error of rejecting a "correct" null hypothesis, where in this case the null hypothesis is that the annotations do not match). But it also shows a low Type II error rate of 5% (the error of accepting a "false" null hypothesis) using a mismatch probability threshold of 0.05 as 29 of the 794 matching annotations were mistakenly identified as dissimilar. The Type II error rate is the more important error statistic with regard to avoiding cascading errors among sequence databases. This is compared to a Type I error rate of 100% and a Type II error rate of 0% for
annotations based on the best E-value (and hence p-value) method where 58 annotations out of 852 hits would have been transferred incorrectly to the query sequence.

A similar approach to the incremental mismatch probability statistic was developed using comparisons of Structural Classification of Proteins (SCOP) protein domains in which a simple empirical approach was used to calculate the significance of an alignment score based on an all-versus-all comparison of the domain database and then curve fitting to the distribution of scores of true negatives (Levitt and Gerstein 1998). This approach expressed the significance of a given alignment score in terms of a p-value, i.e., the chance that an alignment of two randomly selected proteins would obtain this score, and gave values that were in agreement with the differently derived scores from the BLAST and FASTA alignment applications. Aside from improved performance with regard to filtering dubious annotations, an important advantage of the incremental mismatch probability compared to this approach is the wider generality of the mismatch probability statistic because it is based on a larger and more diverse database and can be applied to any database with annotations that are susceptible to the fuzzy matching method. The utility of the incremental mismatch probability method lies also in the fact that the distribution of similar and dissimilar annotations for the potential target only has to be calculated once. The mismatch probability for any future query can then be determined using the E-value of the query to the target.

**Conclusion**

Annotating large new database of ESTs, cDNAs or genomic ORFs cannot be readily done by manual approaches. For example for 10,000 sequences and 5 minutes used to manually
check and annotate each sequence, it would take more than twenty 40-hour work weeks. Thus, most annotation schemes use a computational approach based on sequence similarity. Our findings indicate that the incompleteness and interdependence of databases can significantly affect the accuracy of annotations produced by the most common method - sequence comparison using one or more databases. From these analyses, it is clear that many annotations based on similarities to other DNA sequences will often be missing or incorrect. However, by combining databases, only a few percent of mammalian genes seem to be missing. The other problem is the quality of the target annotation. Assuming the human annotations were correct, one can avoid incorrect annotations by using more than one database, setting a low E-value cutoff, and calculating an incremental mismatch probability for each best hit and putative orthologue. If a query has significant hits in two databases and these two annotations are similar, most likely this annotation is correct (99% for human versus mouse and Swiss-Prot). Although E-values of $10^{-5}$ or less are often used to maximize the percent of annotations, using much lower E-values ($10^{-50}$) yields >80% hits with correct annotation (one that matched the human query). Although with a lower E-value there is an increases chance of a correct annotation, there is also an increase possibility of missing orthologous sequences in one or more target databases. Finally, the incremental mismatch probability provides a statistic for the likelihood of a correct annotation that can be used to significantly reduce the rate of incorrect annotations. Its efficacy in determining incorrect annotations is due to the fact that it is based on a comparison of the target sequence against the entire target database, rather than considering solely the similarity score of the query-target pair. The incremental mismatch probability
approach successfully reduces by half the number of incorrect annotations based on BLAST E-value ‘best hits’. One of the limitations of the incremental mismatch probability approach is its reliance on existing annotations. If a sufficient number of annotations are incorrect, this may undermine or invalidate the incremental mismatch probability approach. Possible reasons for false negatives (the mismatch probability prediction incorrectly indicating a mismatch between the source and target sequences) include:

1) The sequences are not homologous – this is a true type II (false negative) error associated with the algorithm and data set rather than an error of manual annotation.

2) The manual annotation is incorrect because the annotations are incomplete; the sequences are homologous and share multiple functions but only a subset of their functions have been ascribed to the sequences in the two different databases.

3) The manual annotation is incorrect in that the sequences are homologous and share the same function but their function has been incorrectly ascribed in one or both databases.

4) The manual annotation is incorrect/incomplete - the sequences are homologous and share the same function but the description of their function in the two different databases is sufficiently lexically divergent such that even manual annotation results in a mismatch.

Of the above reasons, 2 and 4 appear quite likely given divergence in gene naming and interdependence of databases. For example, the Drosophila gene ‘tiptop’, encodes a protein highly similar to the Drosophila ‘teashirt’ gene, a close paralogue. One of the 1,000 human UniGenes is annotated as ‘Teashirt family zinc finger 3’ and its best hit in the Swiss-Prot
database is a gene called ‘Tiptop protein’. It is conceivable that the UniGene annotation is incorrect, given its imprecision. Given this scenario, combining an incremental mismatch probability prediction with a manual matching or automated fuzzy matching step can serve to flag dubious annotations. One factor militating against reason 3 is the tendency for databases to be annotated using BLAST comparisons such that discordances in annotations between similar sequences would be more likely to be discovered before the distribution of a newly-annotated sequence database.

Possible reasons for false positives (the mismatch probability prediction incorrectly indicating a similar annotation between the source and target sequences) include:

1) The sequences are not in fact homologous – this is a true type I (false positive) error associated with the algorithm and data set rather than an error of manual annotation.

2) The manual annotation is incorrect for one or both sequences and the sequences are in fact homologous and share the same function.

A close examination of the false positives revealed that gene annotations that contain one-word descriptions or synonyms were more likely to be incorrectly predicted as matches when using low match cutoff thresholds. For example, the UniGene sequence BC002748, annotated as ‘transmembrane protein 70’, has its best BLAST hit (E-value: 1.34 x 10^{-45}) against Swiss-Prot sequence P60866, annotated as ‘40S ribosomal protein S20’. The incremental mismatch probability prediction, based on the BLAST hits profile of Swiss-Prot sequence P60866, indicates that they are homologous and share the same function(s). However, it appears unlikely that the same protein could share such different functions. It is
possible that the annotation of one or both sequences is incorrect, or that the use of low E-values to indicate homology has been unsuccessful in this case. In this scenario, the mismatch probability prediction combined with a cursory manual matching or automated fuzzy matching step can serve to flag dubious annotations.

It is worth noting the possibility of a "Type III error", i.e., correctly rejecting the null hypothesis for the wrong reason (Mosteller 1948). In the case of comparison of gene annotations, it is quite conceivable that lexically similar but logically different annotations for the same gene function could be predicted as matches by fuzzy matching.

While a fuzzy match comparison of prospective homologues with similar E-values can also serve to flag dubious annotations among BLAST hits, with regards to its utility in determining the incremental mismatch probability prediction, the high sensitivity but lower specificity of the fuzzy matching heuristic means that the distribution of hits for a particular sequence will tend to over-estimate the number of correct hits. This means that, for example, the associated E-value of a 0.05 incremental mismatch probability will be higher (less negative) than the E-value corresponding to an actual 5% chance of an incorrect annotation.

Nevertheless, it is clear that a relatively unrefined fuzzy match heuristic such as this can accurately detect matching sequence annotations from different databases and thus provide a useful tool for verifying BLAST hits. In addition to ignoring certain ‘stop words’ within annotations, it may be possible to reduce the type II error rate of the fuzzy match by using comprehensive synonym lists. The algorithm’s overall high sensitivity and relatively high specificity regardless of the databases being matched suggests it could be applied to verifying annotations from a wide range of databases. In conjunction with the incremental mismatch
probability, fuzzy matching can also be used to verify the annotations of existing mammalian genomes, and potentially most vertebrate genomes, using BLAST searches against large, relatively complete target databases such as Swiss-Prot.
References


Table 2 - 1 BLAST of 1,000 human UniGenes.

One thousand well-annotated human UniGenes were BLAST against four databases (total numbers of sequences in each database are in parentheses). “Query hits” lists the number of human sequences that had a match with an E-value of less than $10^{-5}$. “Target hits” lists the total number of sequences that match the human sequences in each database. “% of Best hits” lists the percent of human sequences where the match with the lowest E-value came from the specified database.

<table>
<thead>
<tr>
<th>Database</th>
<th>Query hits</th>
<th>% hit</th>
<th>Target hits</th>
<th>% of Best hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss-Prot (188,477)</td>
<td>932</td>
<td>93%</td>
<td>28,822</td>
<td>54%</td>
</tr>
<tr>
<td>UniGene-mouse (64,632)</td>
<td>929</td>
<td>93%</td>
<td>3,836</td>
<td>12%</td>
</tr>
<tr>
<td>UniGene-dog (23,611)</td>
<td>814</td>
<td>81%</td>
<td>1,640</td>
<td>31%</td>
</tr>
<tr>
<td>UniGene-rat (52,183)</td>
<td>930</td>
<td>93%</td>
<td>3,016</td>
<td>13%</td>
</tr>
<tr>
<td>UniGene (140,0426)</td>
<td>967</td>
<td>97%</td>
<td>8,492</td>
<td>50%</td>
</tr>
<tr>
<td>all (328,903)</td>
<td>984</td>
<td>98%</td>
<td>37,314</td>
<td>100%</td>
</tr>
</tbody>
</table>
Table 2 - 2 Number of annotations that are dependent or poor quality.

The annotations of genes that match human-1,000 genes with E-values less than $10^{-5}$ were manually inspected. ‘Dependent’ lists annotations that match the human-1,000 and contain the terms “similar”, predicted’ or ‘hypothetical”. ‘Poor quality’ lists the number of dependent annotations and annotations that contain only a clone number, chromosomal location, or annotation referring to an obscure phenotype.

<table>
<thead>
<tr>
<th>Database</th>
<th>Hits</th>
<th>Dependent</th>
<th>%</th>
<th>Poor quality</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss-Prot</td>
<td>28,820</td>
<td>1,918</td>
<td>6.7%</td>
<td>3,655</td>
<td>12.7%</td>
</tr>
<tr>
<td>UniGene mouse</td>
<td>3,835</td>
<td>469</td>
<td>12.2%</td>
<td>990</td>
<td>25.8%</td>
</tr>
<tr>
<td>UniGene dog</td>
<td>1,639</td>
<td>1,181</td>
<td>72.1%</td>
<td>1,397</td>
<td>85.2%</td>
</tr>
<tr>
<td>UniGene rat</td>
<td>3,015</td>
<td>1,621</td>
<td>53.8%</td>
<td>2,107</td>
<td>69.9%</td>
</tr>
<tr>
<td>All</td>
<td>37,309</td>
<td>5,189</td>
<td>13.9%</td>
<td>8,149</td>
<td>21.8%</td>
</tr>
</tbody>
</table>
Table 2 - 3 Relative E-values of the next best hits for three mammalian UniGenes.

Ratios are calculated as ln(Next hit E-value) / ln(Best hit E-value). The number of hits for the range ratios for the 2nd, 3rd and 4th best hits are given in each column. “%” is the cumulative percentage of annotations with the specified ratio or higher.

<table>
<thead>
<tr>
<th>Ratio</th>
<th># 2nd hits</th>
<th>%</th>
<th># 3rd hits</th>
<th>%</th>
<th># 4th hits</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15-0.2</td>
<td>9</td>
<td>100%</td>
<td>37</td>
<td>100%</td>
<td>78</td>
<td>100%</td>
</tr>
<tr>
<td>0.20-0.25</td>
<td>7</td>
<td>99%</td>
<td>18</td>
<td>96%</td>
<td>47</td>
<td>92%</td>
</tr>
<tr>
<td>0.25-0.3</td>
<td>9</td>
<td>98%</td>
<td>28</td>
<td>94%</td>
<td>34</td>
<td>87%</td>
</tr>
<tr>
<td>0.30-0.35</td>
<td>10</td>
<td>97%</td>
<td>33</td>
<td>91%</td>
<td>46</td>
<td>83%</td>
</tr>
<tr>
<td>0.35-0.4</td>
<td>12</td>
<td>96%</td>
<td>28</td>
<td>88%</td>
<td>51</td>
<td>78%</td>
</tr>
<tr>
<td>0.40-0.45</td>
<td>13</td>
<td>95%</td>
<td>36</td>
<td>85%</td>
<td>59</td>
<td>72%</td>
</tr>
<tr>
<td>0.45-0.5</td>
<td>25</td>
<td>94%</td>
<td>54</td>
<td>81%</td>
<td>75</td>
<td>66%</td>
</tr>
<tr>
<td>0.50-0.55</td>
<td>28</td>
<td>91%</td>
<td>52</td>
<td>75%</td>
<td>58</td>
<td>58%</td>
</tr>
<tr>
<td>0.55-0.6</td>
<td>32</td>
<td>88%</td>
<td>48</td>
<td>70%</td>
<td>45</td>
<td>52%</td>
</tr>
<tr>
<td>0.60-0.65</td>
<td>33</td>
<td>85%</td>
<td>57</td>
<td>65%</td>
<td>65</td>
<td>47%</td>
</tr>
<tr>
<td>0.65-0.7</td>
<td>45</td>
<td>81%</td>
<td>71</td>
<td>58%</td>
<td>74</td>
<td>40%</td>
</tr>
<tr>
<td>0.70-0.75</td>
<td>49</td>
<td>77%</td>
<td>77</td>
<td>51%</td>
<td>72</td>
<td>32%</td>
</tr>
<tr>
<td>0.75-0.8</td>
<td>74</td>
<td>72%</td>
<td>96</td>
<td>43%</td>
<td>74</td>
<td>24%</td>
</tr>
<tr>
<td>0.80-0.85</td>
<td>98</td>
<td>64%</td>
<td>97</td>
<td>33%</td>
<td>54</td>
<td>16%</td>
</tr>
<tr>
<td>0.85-0.9</td>
<td>132</td>
<td>54%</td>
<td>80</td>
<td>22%</td>
<td>50</td>
<td>10%</td>
</tr>
<tr>
<td>0.90-0.95</td>
<td>136</td>
<td>40%</td>
<td>69</td>
<td>14%</td>
<td>31</td>
<td>5%</td>
</tr>
<tr>
<td>0.95-1.0</td>
<td>244</td>
<td>26%</td>
<td>61</td>
<td>6%</td>
<td>16</td>
<td>2%</td>
</tr>
<tr>
<td>Total</td>
<td>956</td>
<td></td>
<td>942</td>
<td></td>
<td>929</td>
<td></td>
</tr>
</tbody>
</table>
Table 2 - 4 Evaluation of FuzzyMatch performance.
The annotations for the best hits in Swiss-Prot and the human 1,000 queries were manually compared to determine true and false positives.

<table>
<thead>
<tr>
<th>Fuzzy cutoff</th>
<th>Total</th>
<th>True positive</th>
<th>True negative</th>
<th>False positive</th>
<th>False negative</th>
<th>Type I error rate</th>
<th>Type II error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>870</td>
<td>362</td>
<td>457</td>
<td>55</td>
<td>12</td>
<td>0.110</td>
<td>0.032</td>
</tr>
<tr>
<td>0.2</td>
<td>870</td>
<td>361</td>
<td>474</td>
<td>38</td>
<td>13</td>
<td>0.076</td>
<td>0.035</td>
</tr>
<tr>
<td>0.3</td>
<td>870</td>
<td>358</td>
<td>483</td>
<td>29</td>
<td>16</td>
<td>0.058</td>
<td>0.043</td>
</tr>
<tr>
<td>0.4</td>
<td>870</td>
<td>352</td>
<td>492</td>
<td>20</td>
<td>22</td>
<td>0.040</td>
<td>0.059</td>
</tr>
<tr>
<td>0.5</td>
<td>870</td>
<td>352</td>
<td>494</td>
<td>18</td>
<td>22</td>
<td>0.036</td>
<td>0.059</td>
</tr>
<tr>
<td>0.6</td>
<td>870</td>
<td>344</td>
<td>494</td>
<td>18</td>
<td>30</td>
<td>0.036</td>
<td>0.081</td>
</tr>
<tr>
<td>0.7</td>
<td>870</td>
<td>339</td>
<td>495</td>
<td>17</td>
<td>35</td>
<td>0.034</td>
<td>0.094</td>
</tr>
<tr>
<td>0.8</td>
<td>870</td>
<td>337</td>
<td>495</td>
<td>17</td>
<td>37</td>
<td>0.034</td>
<td>0.100</td>
</tr>
<tr>
<td>0.9</td>
<td>870</td>
<td>335</td>
<td>496</td>
<td>16</td>
<td>39</td>
<td>0.032</td>
<td>0.105</td>
</tr>
<tr>
<td>1.0</td>
<td>870</td>
<td>329</td>
<td>496</td>
<td>16</td>
<td>45</td>
<td>0.032</td>
<td>0.121</td>
</tr>
</tbody>
</table>
Table 2 - 5 Comparison of Swiss-Prot and Mouse Annotations.

The joint set of target sequences from a BLAST query of the human 1,000 sequences against Swiss-Prot and Mouse UniGene. Annotations are divided into three groups (gray bars): both targets (hits) have similar annotations, targets have dissimilar annotations, and 3 one or both target have non-functional annotations.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sequences</td>
<td>886</td>
<td>100.00%</td>
</tr>
<tr>
<td><strong>Both Target Annotation Similar</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both target annotations are similar to each other and also similar to the query annotation</td>
<td>370</td>
<td>98.90%</td>
</tr>
<tr>
<td>Targets with similar annotations but not similar to human query</td>
<td>4</td>
<td>1.10%</td>
</tr>
<tr>
<td><strong>Targets with dissimilar functional annotations</strong></td>
<td>133</td>
<td>15.00%</td>
</tr>
<tr>
<td>The lowest E-value hit annotation is similar to the query annotation</td>
<td>91</td>
<td>68.40%</td>
</tr>
<tr>
<td>The second E-value hit annotation is similar to the query annotation</td>
<td>11</td>
<td>8.30%</td>
</tr>
<tr>
<td>Both target annotations are dissimilar to the query annotation</td>
<td>31</td>
<td>23.30%</td>
</tr>
<tr>
<td><strong>One or both have non-functional annotations</strong></td>
<td>379</td>
<td>42.80%</td>
</tr>
<tr>
<td>Only one of the targets has non-functional annotations</td>
<td>378</td>
<td>99.70%</td>
</tr>
<tr>
<td>Both target annotations are non-functional</td>
<td>1</td>
<td>0.30%</td>
</tr>
<tr>
<td>Swiss-Prot annotation is non-functional</td>
<td>4</td>
<td>1.10%</td>
</tr>
<tr>
<td>Mouse UniGene annotation is non-functional</td>
<td>376</td>
<td>99.50%</td>
</tr>
<tr>
<td>Only one annotation is functional and it is similar to the query annotation</td>
<td>307</td>
<td>81.20%</td>
</tr>
<tr>
<td>Only one annotation is functional and it is NOT similar to the query annotation</td>
<td>71</td>
<td>18.80%</td>
</tr>
<tr>
<td><strong>Total matches from one or more annotation with best E-value</strong></td>
<td>768</td>
<td>86.70%</td>
</tr>
</tbody>
</table>
Table 2 - 6 Evaluation of the incremental mismatch probability method to validate 'best hit' annotations.

The match/mismatch distribution of BLAST hits for the comparison of the 'best hit' sequence against its own database is first determined. The incremental mismatch probability is defined as the ratio of mismatches to total hits with E-values equal to or less than a given E-value. This is equivalent to the false positive rate for that E-value; the likelihood of reaching a false conclusion (i.e., the wrong annotation) by chance if we assume that all BLAST hits with E-values equal to or less than the given E-value are correct. For each collection sequence, the incremental mismatch probability for its Swiss-Prot 'best hit' was calculated from the BLAST hits of the Swiss-Prot sequence in the Swiss-Prot database. The designation of 'correct' (annotation match) and 'incorrect' (annotation mismatch) hits was determined by fuzzy matching. All predicted matches and mismatches were verified by manual matching of the original human UniGene annotation and the Swiss-Prot 'best hit' annotation.

<table>
<thead>
<tr>
<th>Mismatch cutoff</th>
<th>Total</th>
<th>True positive</th>
<th>True negative</th>
<th>False positive</th>
<th>False negative</th>
<th>Type I error rate</th>
<th>Type II error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>852</td>
<td>743</td>
<td>33</td>
<td>25</td>
<td>51</td>
<td>0.431</td>
<td>0.064</td>
</tr>
<tr>
<td>0.01</td>
<td>852</td>
<td>744</td>
<td>32</td>
<td>26</td>
<td>50</td>
<td>0.448</td>
<td>0.063</td>
</tr>
<tr>
<td>0.05</td>
<td>852</td>
<td>756</td>
<td>29</td>
<td>29</td>
<td>38</td>
<td>0.500</td>
<td>0.048</td>
</tr>
<tr>
<td>0.10</td>
<td>852</td>
<td>763</td>
<td>28</td>
<td>30</td>
<td>31</td>
<td>0.517</td>
<td>0.039</td>
</tr>
<tr>
<td>0.15</td>
<td>852</td>
<td>765</td>
<td>25</td>
<td>33</td>
<td>29</td>
<td>0.569</td>
<td>0.037</td>
</tr>
<tr>
<td>0.20</td>
<td>852</td>
<td>768</td>
<td>24</td>
<td>34</td>
<td>26</td>
<td>0.586</td>
<td>0.033</td>
</tr>
<tr>
<td>0.25</td>
<td>852</td>
<td>772</td>
<td>19</td>
<td>39</td>
<td>22</td>
<td>0.672</td>
<td>0.028</td>
</tr>
<tr>
<td>0.30</td>
<td>852</td>
<td>778</td>
<td>18</td>
<td>40</td>
<td>16</td>
<td>0.690</td>
<td>0.020</td>
</tr>
<tr>
<td>0.40</td>
<td>852</td>
<td>781</td>
<td>13</td>
<td>45</td>
<td>13</td>
<td>0.776</td>
<td>0.016</td>
</tr>
<tr>
<td>0.50</td>
<td>852</td>
<td>788</td>
<td>7</td>
<td>51</td>
<td>6</td>
<td>0.879</td>
<td>0.008</td>
</tr>
<tr>
<td>0.75</td>
<td>852</td>
<td>791</td>
<td>1</td>
<td>57</td>
<td>3</td>
<td>0.983</td>
<td>0.004</td>
</tr>
<tr>
<td>1.00</td>
<td>852</td>
<td>794</td>
<td>0</td>
<td>58</td>
<td>0</td>
<td>1.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure 2 - 1 Number of significant hits for 1,000 human UniGenes.

The number of sequences that had one or more significant BLAST hits for E-values of less than $10^{-5}$ (blue-diagonal), $10^{-50}$ (gray stipules) and $10^{-100}$ (red strips is shown. Three BLAST searches are depicted: A) is for the 932 human UniGenes that had matches in Swiss-Prot; B) is for the 967 human UniGenes that had matches in one of the three mammalian UniGene databases and C) is for 984 sequences that were found in either Swiss-Prot or the mammalian databases.
Fuzzy match flowchart

Example
Choriogonadotropin
Chorionic gonadotropin

Choriogonadotropin  gonadotropin  Score = 1

Choriogonadotropin  Chorionic

Choriogonadotropin  Chorionic

Choriogonadotropin  Chorionic

Choriogonadotropin  Chorionic  Score = 4/7

Total score = (1 + 4/7) / 2 = 11/14 = 0.786

Figure 2 - 2 FuzzyMatch Algorithm.
For any two input annotations, a fuzzy match score from 0 to 1 is calculated, representing the possibility that the two annotations are the same.
Figure 2 - 3 Matched and Mismatched Annotations.

A. The number of hits with matching or mismatching annotations relative to the E-value for BLAST hits of human UniGene sequences against Swiss-Prot. The numbers above the bars represent the fraction of total annotations where the human query annotation matches the Swiss-Prot hit annotation. For E-values of $10^{-201}$ to $10^{-300}$ there are 174 matches and eight mismatches. B. The proportion of matches and mismatches for hits with E-values < $10^{-100}$. The numbers above bars are the number of queries with this number of hits. (For example, there are 86 queries with one hit at an E-value of $10^{-100}$.)
Figure 2 - 4 Self-BLAST of Swiss-Prot.

This graph shows the distribution of the billions of matches and mismatches for 188,447 Swiss-Prot queries when BLAST against the Swiss-Prot database. The numbers above the histogram bars are the proportion of matching annotations.
Figure 2 - 5 Incremental mismatch probability for human UniGene S15967514 (Cytochrome c oxidase subunit 8A).

Cumulative sum of the number of BLAST hits (orange triangles) among human UniGenes relative to the E-value for these hits. Incremental mismatch probability (blue line) with mismatched annotations shown as red triangles. A sharp rise in the incremental mismatch probability (decreased significance) is associated with a hit that has a mismatched annotation.
Chapter 3: InterPro Motif Annotation Trees (IMAT)
Abstract

The use of sequence similarity search tools to annotate unknown sequences can be problematic due to missing orthologues in target databases, multiple hits in different target databases with divergent annotations, cascading annotation errors and other issues. The InterPro Motif Annotation Trees (IMAT) resource provides an alternative sequence annotation source without these drawbacks by using protein motifs, which are found in biological sequences as frequently as BLAST hits at commonly used E-value thresholds.

InterPro motif names comprise a structured vocabulary that can be used as attributes in machine learning classifiers. Decision Tree and Naïve Bayes machine learning models based on InterPro motif signatures were evaluated as predictors of protein function with a diverse assortment of orthologue and EST databases. Decision Trees showed superior performance compared to Naïve Bayes, indicating that the order of motifs is an important functional indicator. Optimized attribute construction parameters were used to construct Decision Trees for all unique Swiss-Prot functional annotations and these were comprehensively evaluated as function predictors for randomly selected Swiss-Prot sequences.

The IMAT method uses the unique order of protein motifs in the largest possible data source - all known proteins in prokaryotes, eukaryotes and archaea – to identify the precise function of randomly selected proteins with a relatively high degree of accuracy. IMAT can be used to filter annotations, annotate unknown sequences and as generalized models representing the dynamic between protein motif signatures and protein function.
Introduction

There is an urgent need for alternative annotation methods to check the consistency of BLAST annotations and detect potential annotation errors using independent sources of evidence in a high-throughput fashion (Andorf, et al. 2007). Protein motifs - amino acid patterns found in proteins with the same function (Bork and Koonin 1996) - are one such source. Because motifs represent highly conserved regions of proteins sharing the same function, they can provide clues to a protein’s role even if it does not share large regions of similar sequence with any known protein (Nevill-Manning, et al. 1998).

The relationship between the building blocks of proteins and their function has been studied at different levels of detail. On the larger scale, the fusion of two protein domains to create a new function is thought to be an important avenue for the evolution of proteins. The order of fused domains is conserved over the course of evolution (Apic, et al. 2001) with almost absolute preservation of the N to C-terminal orientation of combined domain pairs (Pemberton 2006). Whilst the most common protein domain size is approximately 150 amino acids (Shen, et al. 2005) within a well-characterized distribution of domain sizes (Wheelan, et al. 2000), on the finer scale, the average length of InterPro motifs (Mulder, et al. 2003) (Mulder and Apweiler 2008) in the UniProt/Swiss-Prot database (Bairoch, et al. 2004) (Schneider, et al. 2005) is 49 amino acids (147 nucleotides). The order of protein motifs is an important determinant of the function of proteins (Bashton and Chothia 2002). The evolutionary conservation of motif and motif order can be used for annotations. These
approaches are appropriate for the large-scale analysis of protein function but work best with relatively structured vocabularies such as motif databases, where each motif name is uniquely mapped to a particular conserved pattern. Protein motif signatures have previously been used for sequence annotation, notably in Xanthippe, a tool for filtering out incorrect annotations for the UniProt database (Wieser, et al. 2004) and in Spearmint, which automatically generates Decision Tree rules to assign one of ~800 Swiss-Prot keywords to unannotated TrEMBL sequences (Kretschmann, et al. 2001). Sequence-pair motif content similarity scores have also been used as a kernel for Support Vector Machines to detect remote homology in cases of low sequence similarity (Ben-Hur and Brutlag 2003).

Alongside machine learning approaches, Rulebase generates manually curated rules to group proteins by their InterPro motifs and extract the common annotation in each group (Biswas, et al. 2002). Similarly, manually curated rules based on full-length sequence similarity and common domain architecture have also been used to classify proteins into protein families (Wu, et al. 2004).

A novel approach for annotation by using InterPro Motif Annotation Trees (IMAT) methodology was developed in three stages: design, validation and implementation. The design stage involved characterization of the occurrence and distribution of InterPro motifs with respect to unique Swiss-Prot functional annotations (Boeckmann, et al. 2005) and the development of a method of transforming protein motifs into sequence-specific attributes suitable for machine learning. The validation stage consisted of two parts: 1) evaluating the position-specific component of the motif signature by comparing the performance of
Decision Trees and Naïve Bayes classifiers on a variety of orthologue collections, and 2) optimizing the efficacy of motif-based Decision Trees for functional prediction using multiple data sets. The implementation stage involved the construction of Decision Trees for all unique Swiss-Prot annotations using the optimal parameters derived from the validation stage.

Methods

2.1 Motif distributions and correlations

For the first part of the design stage, InterPro database motifs were used because they provide a large number of motifs with a structured vocabulary (Apweiler, et al. 2000). Pearson correlation coefficients were calculated for all possible pairs of InterPro motif databases for Swiss-Prot, three orthologue databases and the random sample of 1,000 human UniGene described below. The second part of the design stage involved detailed investigation of the relationship between motif names, and between motif names and protein annotations. The general abundance of InterPro motifs compared to BLAST hits at an E-value cutoff of $10^{-5}$ was also evaluated using 1,000 randomly selected human UniGenes.

2.2 Validating Decision Trees

The first part of the validation stage involved validating the hypothesis that the sequence of motifs is a key factor in protein function. Decision Tree and Naïve Bayes models were tested using 10-fold cross validation, or less depending on the number of identified orthologues and non-orthologues in each training/test set, with four different orthologue sequence collections.
In the second part, the optimal parameters for Decision Trees were derived based on the overall performance of Decision Tree models on these data sets.

2.3 Human UniGene orthologues

One thousand randomly-selected human UniGene sequences with well-defined gene functions not containing the words ‘putative’ or ‘hypothetical’ were chosen, along with sequences with similar combinations of protein motifs in the Swiss-Prot database (release 51.1, 141,991 sequences), to form training/test sets of orthologous and non-orthologous proteins. The best BLAST hits of these 1,000 human UniGenes among Swiss-Prot and the mouse, dog and rat UniGene collections were used for an additional analysis to determine whether Decision Trees could be successfully constructed for these BLAST hits and if Decision Trees based on the human UniGenes could be used to identify these BLASTs hits.

2.4 CluSTr orthologues

Two datasets were used from CluSTr (Petryszak, et al. 2005), one of the largest existing orthologue databases cluster orthologues: 1) Homo sapiens (human) and Mus musculus (mouse) orthologues and, 2) Plasmodium falciparum (malaria) and Anopheles gambiae (mosquito) orthologues. The human-mouse collection contained 23,260 sequences in 11,630 orthologue pairs. The malaria-mosquito collection consisted of a total 3,496 sequences in 1,375 orthologue groups of two or more sequences. Of these groups only 1,183 had sufficient annotation to allow a lexical search for matches among Swiss-Prot sequences sharing the same protein motifs to determine whether the Swiss-Prot sequences also shared the same function as the orthologue group. The Swiss-Prot sequences were classified as ‘1’
if they had same function, ‘0’ if had a different function. The CluSTr sequences were downloaded from the Integr8 website (EMBL-EBI 2007): http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do.

2.5 Eukaryotic orthologues (KOG)

The Eukaryotic Collections of Orthologous Genes (KOG) consists of 4,852 clusters of orthologues, including 60,759 proteins from a total of 112,920 analyzed eukaryotic gene products (NCBI 2007). KOG includes proteins from 7 eukaryotic genomes: three animals (the nematode Caenorhabditis elegans, the fruit fly Drosophila melanogaster and Homo sapiens), one plant, Arabidopsis thaliana, two fungi (Saccharomyces cerevisiae and Schizosaccharomyces pombe), and the intracellular microsporidian parasite Encephalitozoon cuniculi. The KOG orthologue groups were larger compared to the CluSTr and UniGene orthologue groups (Tatusov, et al. 2003), varying in size from two to 136 putative orthologues from different species per orthologue group.

2.6 Orthologue training/test set selection

The four sequence collections (three orthologue databases and a sample of 1,000 expressed sequence tags (ESTs) from the UniGene database) contained groups of two or more orthologous sequences (one sequence per group in the case of the UniGene sample). In order to achieve a sufficiently large set of ‘correct’ sequences (i.e., proteins sharing the same function) and ‘incorrect’ sequences (i.e., proteins with a different function) with a relatively homogenous annotation format, Swiss-Prot proteins that shared the most InterPro motifs in common with each orthologue group were used to create a test/training set. For each putative
orthologue group, two sequences were selected as a ‘seed’ set. All of the unique InterPro motifs found in the seed were identified and the Swiss-Prot sequences that most shared these motifs were selected to form the training/test set.

The number of seed sequences was limited to two sequences because both the Naïve Bayes and Decision Tree classifiers performed very poorly on the KOG data set when no seed limit was used. This poor classifier performance on KOG sequences in the absence of a seed limit was likely due to the presence of sequences from more than one functional group (i.e., proteins possessing different unique protein motif signatures) within the same KOG groups.

The selected Swiss-Prot training/test set sequences were added to the seed sequences and classified as ‘1’ if their annotations matched any of the annotations for the sequences in the seed set using fuzzy matching. The fuzzy matching algorithm had a 3.6% Type I error rate and 5.9% Type II error rate for matching UniGene and Swiss-Prot annotations using a fuzzy match score threshold of 0.5 (Table 3 - 1). Non-matching sequences were classified as ‘0’, i.e., they were considered to have a different function to that of the seed set. Equal numbers of ‘1’ and ‘0’-classified sequences were selected (20 sequences each, total 40) to meet the requirement of balanced sets to avoid bias using Naïve Bayes models.

2.7 Swiss-Prot tree training/test set selection

For implementation stage, Decision Trees were constructed for all unique Swiss-Prot annotations. For each unique annotation, at least one and up to 10 sequences sharing the annotation were used to populate the seed set, and were added to the training/test set and
classified as ‘1’. An additional number of Swiss-Prot sequences with different annotations but which most shared the protein motifs of the seed set were classified as ‘0’ outcomes and added to make 40 sequences in total for each training/test set. Each training/test set was thus composed of one to ten ‘1’-classified sequences sharing the same unique Swiss-Prot functional annotation and 30 to 39 ‘0’-classified sequences not sharing this Swiss-Prot annotation but, as much as possible, sharing the protein motifs found in the ‘1’-classified sequences.

A training/test set was composed for each of the 68,686 unique Swiss-Prot annotations (out of a total 88,410 unique annotations) in the 54.6 release that had InterPro hits (i.e., 276,411 out of 290,484 sequences). Decision Trees were constructed these functional groups using a fixed domain size of 300 amino acids. The Decision Tree analysis was not performed in the trivial case where a protein motif only occurs in one functional group and so can be used as a unique identifier.

### 2.8 Motif categorization

Motif categorization for the Naïve Bayes method involved listing the counts of all the motifs found in each orthologue group and inputting them directly into the Naïve Bayes model. For the Decision Tree method, we used discrete domains or segments of the multiple alignment of the orthologue sequences as attributes and the longest InterPro name among the motifs found in each domain as the attribute values. The domain boundaries were defined using three different methods: 1) InterPro motif boundaries with intervening regions allocated equally to neighboring motif groups, 2) fixed-size domains, and 3) a combination of both
where the intervening regions were defined as separate domains of fixed size.

First, the complete training/test set for each orthologue group, containing orthologue and Swiss-Prot sequences, was aligned using ClustalW (Thompson, et al. 1994) (Figure 3 - 1).

Next, the InterProScan hits of each orthologue sequence in the multiple alignment were collected, their residue positions were adjusted by their positions in the multiple alignment, and they were ordered by motif start point from most upstream (5’) to most downstream (3’). The domains were then assembled as follows:

a. The first domain is designated as encompassing the most upstream motif. Beginning at the next most upstream motif, downstream motifs are added as new domains to the list of domains if they do not overlap with any of the domains in the domain list. Otherwise, the motif is added to the domain.

b. If a motif has been added to a domain, the recipient domain is then split into two domains if a different motif from the same InterPro database as the added motif’s database (e.g., PFAM, SMART) already exists in the domain.

The sequence attributes were created based on the InterPro motifs that were found within the domain boundaries, as shown in Figure 3 - 2. Figure 3 - 3 shows a Decision Tree model built using InterPro motifs and the overall application workflow is described in Figure 3 - 4.

2.9 Swiss-Prot Decision Tree evaluation

In the implementation stage, a two-step approach was used for evaluating the predictive
performance of Swiss-Prot Decision Trees. First, prospective Decision Trees were filtered based on the number of consecutive motif pairs that the unknown shared with the original InterPro domain upon which the Decision Tree was based (i.e., the number of motifs in common in the case where the sequence had only one domain). Second, the lookup list of Decision Trees was input into the second stage where the trees for the top-ranking five functional groups were tested against the unknown sequence. This was accomplished by comparing the protein’s motif signature against each Decision Tree and obtaining a prediction (i.e., ‘belongs/does not belong’ to the functional group). For each prediction, an associated probability value was also available based on the actual performance of the Decision Tree during the N-fold cross validation step.

2.10 InterProScan for motif finding

In order to provide as complete an analysis as possible, all of the public InterPro databases were used for model creation, with the exception of SUPERFAMILY due to its relatively long processing time and high correlation with other InterPro databases. For each orthologue collection and the UniGene collection, we used InterProScan to search for Interpro hits (Zdobnov and Apweiler 2001). InterProScan is a Perl 5 software application developed to run on Unix-compatible operating systems (e.g., Unix, Linux, Mac OSX) to search InterPro databases (http://www.ebi.ac.uk/interpro/user_manual.html). InterProScan was run using the Sun Grid Engine job scheduling application on a dedicated 32-CPU Apple Xserve cluster. The InterPro hits for Swiss-Prot were taken from the match_complete.xml (InterPro version 4.3).
Results

3.1 Motif distributions and correlations

Comparing BLAST and protein motif search demonstrated that a particular protein is as likely to contain a protein motif as it is to have a BLAST hit at commonly used E-value cutoffs. A random sample of 1,000 human UniGenes, approximately 97.3% had InterPro hits while 98.3% had BLAST hits against either Swiss-Prot or the UniGene mouse, dog and rat databases with an E-value cutoff of $10^{-5}$. However, InterPro motifs from different databases found in the same sequence region may have different names, many of which are dissimilar to the functional description of the protein they are found in. As a result, the name of the motif cannot be reliably and meaningfully used to describe the protein’s function.

Of the 24,988 unique InterPro motifs found in the Swiss-Prot collection, only 11% appear in only one functionally distinct protein (Figure 3 - 5). A further 27% are found in two to five functionally distinct proteins and 20% are found in more than 50 different proteins. Since identification of protein function using unique motifs markers is only possible in 11% of cases, for the other 89% there is a need for a methodology which uses two or more motifs – the motif signature - to identify proteins.

The highest average correlation of occurrence of motifs across the four different sequence collections (three orthologue collections and a human UniGene sample) was 0.383 and correlation was partially related to the use of similar methodologies for motif construction. The maximum correlation of occurrence of motifs among the Swiss-Prot proteins was 0.493
(Table 3 - 2) and the ranking of motif correlations among Swiss-Prot proteins generally concurred with that of the abovementioned sequence collections (Table 3 - 3).

3.2 Decision Trees outperform Naïve Bayes

The performance of Decision Trees was compared to that of Naïve Bayes models to determine the importance of the particular order of motifs in determining protein function. Decision Trees are position-specific with relation to the sequence of motifs while Naïve Bayes models simply calculate the frequency of the different motifs found in proteins sharing the same functional annotation. Decision Tree models had maximum correct prediction scores ranging from 75% to 86%, compared to maximum scores of 74% to 76% for the Naïve Bayes model (Table 3 - 4). This outcome supports the hypothesis that the sequence of motifs is a key determinant in protein function. On average, a fixed domain size of 300 amino acids allowed the construction of Decision Trees that were better classifiers than using the other two domain categorization methods: InterPro hit boundaries, and a combination of both InterPro boundaries and fixed size domains. The average prediction accuracies were calculated by 10-or-less cross validation.

3.3 Large, fixed length attributes are optimal

Among the Decision Tree models, the larger, sized-based domains consistently performed better than the InterPro motif-based domains and the InterPro-size combination domains. Decision Trees constructed with a 300bp domain width were most effective at looking up the sequences upon which they were based in the first step of testing - the number of shared
consecutive motif pairs – for 78.5% of the 1,000 human UniGenes. Of the other 21.5%, domains created using a combination of InterPro boundaries and fixed domain sizes of 150 bp and 100 bp scored higher numbers of shared motif pairs than Decision Trees constructed with a 300 bp domain width. Because of their overall better performance, Decision Trees based on a fixed domain width of 300 amino acids were constructed for all 1,000 human UniGene and tested using the 855 top BLAST hits of these sequences for which attributes could be constructed. Of the 855 BLAST hits, 617 were correctly identified with the Decision Tree of the corresponding human UniGene sequence and 28% (238 BLAST hit sequences) were not correctly identified using Decision Trees with this domain size. Of the total 983 BLAST hits, 43 were verified manually as incorrect hits based on comparison of the original human UniGene annotation with the BLAST hit annotation (Table 3 - 5). Of these incorrect hits, 29% (11 out of 38 BLAST hit sequences) were not correctly identified 38 by the UniGene trees.

### 3.4 Swiss-Prot trees predict function

The performance of Swiss-Prot motif-based Decision Trees is shown in (Table 3 - 6). In total, Decision Trees were able to be constructed for 53,895 unique Swiss-Prot annotations out of a total of 68,686 unique Swiss-Prot annotations. Trees were not constructed for functional groups that do not have motifs in common with any other group. Among the 53,895 unique Swiss-Prot protein annotations, 82% (44,345 groups) traced back to the original annotation based on a lookup of consecutive motif pairs compiled from up to 10 members of the functional group. For some protein functional groups, randomly selected
members of the group did not achieve predictions because their motifs were different to those of the seed proteins for their group. In 65% of protein groups, members had tree predictions for their original annotation group, and 88% (57% of all groups) were predicted as matches. A total of 14% protein groups were predicted as belonging solely to the original functional group and 43% that were also predicted as belonging to at least one other functional group.

A subsample of 500 bp fragments from the 5-prime end of the first sample had 93% successful lookups to the original functional group. The 500 bp fragments had correct prediction and unique correct prediction rates of 78% and 41% respectively. Manual inspection of the cases where 300-amino acid domain-based Decision Trees were unable to determine between label ‘1’ and label ‘0’ functional groups indicated that this was due to insufficient granularity; Decision Trees based on smaller fixed domain widths would have resulted in correct assignments.

**Conclusion**

Correlations between motif databases were generally low, which supports their assumed independence based on the different motif construction methods (Decision Trees, Hidden Markov Models, Support Vector Machines, Neural Nets, manual annotation). This also supports the inclusion of most InterPro databases as input for machine learning as they capture different signals from the underlying biological data. However, motif databases with generally high correlations to several other databases, such as SMART, could be excluded from the analysis in order to reduce computational costs.
The relative efficiency of Decision Trees based on motif signatures from genes of more than one organism showed that classifiers with prediction accuracies as high as 91% of total self predictions can be used as generalized models of protein function. The level of accuracy of machine learning methods may also be taken as a reflection of the accuracy of the orthologue prediction algorithms used to construct the databases. Both classifiers performed relatively poorly on the KOG data set when there was no limit on the number of seed sequences from each KOG orthologue group, which suggests that KOG orthologue groups may contain more than one actual functional group. This was not the case with the CluSTr data sets, which suggests their orthologue group predictions may be more accurate than those of KOG.

Aside from verifying the accuracy of existing orthologue datasets, InterPro Motif Annotation Trees (IMAT) can be used to determine in a high-throughput manner to predict the function of a proteins or protein fragments. One advantage of Decision Tree classifiers is that the tree can be traversed from any node, ignoring any preceding nodes where the sequence is missing. The accuracy of a tree prediction can be calculated as the percentage of correct classifications it makes using the training/test data set. Prediction accuracies averaged over the actual sequences used in all N validation steps can be used as a measure of the accuracy of an individual prediction. The IMAT approach also provides a novel solution to a particular problem of using biological data for machine learning: meaningful categorization of the biological data for input into efficient machine learning methods, and ease of interpretation of the results. By using motif names as inputs and outputs, IMAT facilitates insights into the relationship between protein motifs and protein function as IMAT trees are
based on the motif signatures of almost all known proteins in known prokaryotes, eukaryotes and archaea. Despite this heterogeneity of sources, the InterPro motif-based Decision Trees can identify the precise function of a protein and its more general functional group with a relatively high degree of accuracy using only one set of motif attribute construction parameters (300 bp fixed length). The UniGene Decision Tree performance results suggest that shorter attribute lengths are optimal in approximately 20% of proteins, suggesting there is room for improvement in some protein functional groups given sufficient attribute granularity.

Interestingly, the fraction of tree predictions was much higher for 5-prime fragments of proteins drawn from this set, although the prediction performance was much lower. This indicates that, for short sequence fragments, the occurrence of any prediction may be an effective means of annotating protein fragments such as expressed sequence tags (ESTs). Increased tree predictions in the 5-prime portion of proteins may be due to the presence of signal peptides, protein folding determinants and other function-related features that are exclusive to this region.

In summation, this study indicates that machine learning analysis using motif signatures of Swiss-Prot database sequences can be used to construct easily interpretable models of protein function that uniquely and effectively identify Swiss-Prot functional annotations. The current performance may improve significantly by selecting the best performing Decision Trees for each unique orthologue group from a wide range of domain sizes and domain construction methods. An important advantage of the IMAT application of machine learning
is the ability to integrate various kinds of data, such as sequence bias, BLAST hits or other position-specific markers to improve annotation accuracy. We plan to provide the IMAT library of Decision Tree classifiers as an online annotation resource for the genomics community.
References


Apweiler, R. et al., InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16 (12), 1145-1150 (2000).


Kretschmann, E., Fleischmann, W., & Apweiler, R., Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* 17 (10), 920-926 (2001).


Thompson, J.D., Higgins, D.G., & Gibson, T.J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence

Table 3 - Evaluation of fuzzy matching for human 1,000 and Swiss-Prot annotations.

The similarity of annotations of 870 human UniGenes and their best Swiss-Prot BLAST hits was predicted using fuzzy matching. The accuracy of the fuzzy match was determined by manual inspection.

<table>
<thead>
<tr>
<th>Fuzzy cutoff</th>
<th>Total</th>
<th>True positive</th>
<th>True negative</th>
<th>False positive</th>
<th>False negative</th>
<th>Type I error rate</th>
<th>Type II error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>870</td>
<td>362</td>
<td>457</td>
<td>55</td>
<td>12</td>
<td>0.110</td>
<td>0.032</td>
</tr>
<tr>
<td>0.2</td>
<td>870</td>
<td>361</td>
<td>474</td>
<td>38</td>
<td>13</td>
<td>0.076</td>
<td>0.035</td>
</tr>
<tr>
<td>0.3</td>
<td>870</td>
<td>358</td>
<td>483</td>
<td>29</td>
<td>16</td>
<td>0.058</td>
<td>0.043</td>
</tr>
<tr>
<td>0.4</td>
<td>870</td>
<td>352</td>
<td>492</td>
<td>20</td>
<td>22</td>
<td>0.040</td>
<td>0.059</td>
</tr>
<tr>
<td>0.5</td>
<td>870</td>
<td>352</td>
<td>494</td>
<td>18</td>
<td>22</td>
<td>0.036</td>
<td>0.059</td>
</tr>
<tr>
<td>0.6</td>
<td>870</td>
<td>344</td>
<td>494</td>
<td>18</td>
<td>30</td>
<td>0.036</td>
<td>0.081</td>
</tr>
<tr>
<td>0.7</td>
<td>870</td>
<td>339</td>
<td>495</td>
<td>17</td>
<td>35</td>
<td>0.034</td>
<td>0.094</td>
</tr>
<tr>
<td>0.8</td>
<td>870</td>
<td>337</td>
<td>495</td>
<td>17</td>
<td>37</td>
<td>0.034</td>
<td>0.100</td>
</tr>
<tr>
<td>0.9</td>
<td>870</td>
<td>335</td>
<td>496</td>
<td>16</td>
<td>39</td>
<td>0.032</td>
<td>0.105</td>
</tr>
<tr>
<td>1.0</td>
<td>870</td>
<td>329</td>
<td>496</td>
<td>16</td>
<td>45</td>
<td>0.032</td>
<td>0.121</td>
</tr>
</tbody>
</table>
Table 3 - 2 Pairwise correlations of protein motifs from different InterPro databases among three orthologue databases and an expressed sequence tags (EST) collection.

This table shows the correlation of occurrence of protein motifs in all 55 pairwise combinations of hits of InterPro databases (GENE3D, PANTHER, PFAM, PIR, PRINTS, PRODOM, PROFILE, PROSITE, SEG, SMART and TIGRFAMs) was calculated for sequences in four sequence databases: 1,000 randomly selected human UniGenes, the 172 orthologue groups of the Eukaryotic Orthologous Groups of Proteins (KOG) (the first 2 sequences in each orthologue group), and the ChuSTr orthologue predictions for *Plasmodium falciparum* (malaria) and *Anopheles gambiae* (mosquito), and *Homo sapiens* (human) and *Mus musculus* (mouse). A high correlation value indicates a common co-occurrence of motifs from both databases in the same sequence. A negative correlation value indicates the degree to which motifs of either database tend not to occur if a motif from the other database is present in the sequence (top 10 highest average correlations are shaded gray).

<table>
<thead>
<tr>
<th>InterPro Database Pairs</th>
<th>H.sap UniGene</th>
<th>P.fal-A.gam</th>
<th>H.sap-M.mus</th>
<th>KOG</th>
<th>Average correlation</th>
<th>Average correlation ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.572</td>
<td>0.116</td>
<td>0.301</td>
<td>0.241</td>
<td>0.168</td>
<td>11</td>
</tr>
<tr>
<td>GENE3D/PFAM</td>
<td>0.491</td>
<td>0.169</td>
<td>0.204</td>
<td>0.311</td>
<td>0.246</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PFAM</td>
<td>0.106</td>
<td>0.257</td>
<td>0.019</td>
<td>0.013</td>
<td>0.007</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROSITE</td>
<td>0.406</td>
<td>0.120</td>
<td>0.236</td>
<td>0.246</td>
<td>0.246</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/SEG</td>
<td>0.348</td>
<td>0.108</td>
<td>0.186</td>
<td>0.138</td>
<td>0.186</td>
<td>10</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.322</td>
<td>0.326</td>
<td>0.451</td>
<td>0.491</td>
<td>0.311</td>
<td>7</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.149</td>
<td>0.186</td>
<td>0.296</td>
<td>0.247</td>
<td>0.186</td>
<td>10</td>
</tr>
<tr>
<td>GENE3D/SEG</td>
<td>0.019</td>
<td>0.070</td>
<td>0.039</td>
<td>0.013</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.614</td>
<td>0.175</td>
<td>0.181</td>
<td>0.143</td>
<td>0.245</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.182</td>
<td>0.199</td>
<td>0.069</td>
<td>0.056</td>
<td>0.019</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.162</td>
<td>0.211</td>
<td>0.103</td>
<td>0.080</td>
<td>0.011</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.149</td>
<td>0.186</td>
<td>0.296</td>
<td>0.247</td>
<td>0.186</td>
<td>10</td>
</tr>
<tr>
<td>GENE3D/SEG</td>
<td>0.019</td>
<td>0.070</td>
<td>0.039</td>
<td>0.013</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.554</td>
<td>0.199</td>
<td>0.181</td>
<td>0.143</td>
<td>0.245</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.182</td>
<td>0.199</td>
<td>0.069</td>
<td>0.056</td>
<td>0.019</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.162</td>
<td>0.211</td>
<td>0.103</td>
<td>0.080</td>
<td>0.011</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.149</td>
<td>0.186</td>
<td>0.296</td>
<td>0.247</td>
<td>0.186</td>
<td>10</td>
</tr>
<tr>
<td>GENE3D/SEG</td>
<td>0.019</td>
<td>0.070</td>
<td>0.039</td>
<td>0.013</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.554</td>
<td>0.199</td>
<td>0.181</td>
<td>0.143</td>
<td>0.245</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.182</td>
<td>0.199</td>
<td>0.069</td>
<td>0.056</td>
<td>0.019</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.162</td>
<td>0.211</td>
<td>0.103</td>
<td>0.080</td>
<td>0.011</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.149</td>
<td>0.186</td>
<td>0.296</td>
<td>0.247</td>
<td>0.186</td>
<td>10</td>
</tr>
<tr>
<td>GENE3D/SEG</td>
<td>0.019</td>
<td>0.070</td>
<td>0.039</td>
<td>0.013</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.554</td>
<td>0.199</td>
<td>0.181</td>
<td>0.143</td>
<td>0.245</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D/PANTHER</td>
<td>0.182</td>
<td>0.199</td>
<td>0.069</td>
<td>0.056</td>
<td>0.019</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.162</td>
<td>0.211</td>
<td>0.103</td>
<td>0.080</td>
<td>0.011</td>
<td>9</td>
</tr>
<tr>
<td>GENE3D/PROFILE</td>
<td>0.149</td>
<td>0.186</td>
<td>0.296</td>
<td>0.247</td>
<td>0.186</td>
<td>10</td>
</tr>
<tr>
<td>GENE3D/SEG</td>
<td>0.019</td>
<td>0.070</td>
<td>0.039</td>
<td>0.013</td>
<td>0.003</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 3 - 3 Pairwise correlations of protein motifs from different InterPro databases for all SwissProt sequences with InterPro hits (total 276,372).

This table shows the correlation of occurrence of protein motifs in all 55 pairwise combinations of hits of InterPro databases (GENE3D, PANTHER, PFAM, PIR, PRINTS, PRODOM, PROFILE, PROSITE, SEG, SMART and TIGRFAMs) for all sequences in the SwissProt database. A high correlation value indicates a common co-occurrence of motifs of both databases in the same sequence. A negative correlation value indicates the degree to which motifs of either database tend not to occur if a motif from the other database is present in the sequence. The top 10 highest average correlations are shaded gray. This ranking concurs generally with the ranking in Table 3 - 2. The highest correlation is 0.493 between PROFILE motifs and SMART motifs, followed by 0.341 for GENE3D and PANTHER motifs. SMART motifs are also highly correlated with PANTHER and GENE3D motifs.

<table>
<thead>
<tr>
<th>InterPro Database Pairs</th>
<th>Pairwise correlation</th>
<th>Correlation ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE3D PANTHER</td>
<td>0.341</td>
<td>2</td>
</tr>
<tr>
<td>GENE3D PFAM</td>
<td>0.239</td>
<td>7</td>
</tr>
<tr>
<td>GENE3D PRINTS</td>
<td>0.248</td>
<td>6</td>
</tr>
<tr>
<td>GENE3D PRODOM</td>
<td>0.076</td>
<td>24</td>
</tr>
<tr>
<td>GENE3D PROFILE</td>
<td>0.189</td>
<td>11</td>
</tr>
<tr>
<td>GENE3D PROSITE</td>
<td>0.327</td>
<td>3</td>
</tr>
<tr>
<td>GENE3D SMART</td>
<td>0.229</td>
<td>8</td>
</tr>
<tr>
<td>GENE3D TIGRFAMs</td>
<td>0.165</td>
<td>13</td>
</tr>
<tr>
<td>PANTHER PFAM</td>
<td>0.121</td>
<td>21</td>
</tr>
<tr>
<td>PANTHER PRINTS</td>
<td>0.207</td>
<td>10</td>
</tr>
<tr>
<td>PANTHER PRODOM</td>
<td>-0.016</td>
<td>33</td>
</tr>
<tr>
<td>PANTHER PROFILE</td>
<td>0.187</td>
<td>12</td>
</tr>
<tr>
<td>PANTHER PROSITE</td>
<td>0.319</td>
<td>5</td>
</tr>
<tr>
<td>PANTHER SMART</td>
<td>0.214</td>
<td>9</td>
</tr>
<tr>
<td>PANTHER TIGRFAMs</td>
<td>0.030</td>
<td>30</td>
</tr>
<tr>
<td>PFAM PRINTS</td>
<td>0.127</td>
<td>20</td>
</tr>
<tr>
<td>PFAM PRODOM</td>
<td>0.004</td>
<td>31</td>
</tr>
<tr>
<td>PFAM PROFILE</td>
<td>0.053</td>
<td>26</td>
</tr>
<tr>
<td>PFAM PROSITE</td>
<td>0.140</td>
<td>17</td>
</tr>
<tr>
<td>PFAM SMART</td>
<td>0.084</td>
<td>23</td>
</tr>
<tr>
<td>PFAM TIGRFAMs</td>
<td>0.154</td>
<td>16</td>
</tr>
<tr>
<td>PRINTS PRODOM</td>
<td>0.053</td>
<td>27</td>
</tr>
<tr>
<td>PRINTS PROFILE</td>
<td>0.108</td>
<td>22</td>
</tr>
<tr>
<td>PRINTS PROSITE</td>
<td>0.326</td>
<td>4</td>
</tr>
<tr>
<td>PRINTS SMART</td>
<td>0.132</td>
<td>19</td>
</tr>
<tr>
<td>PRINTS TIGRFAMs</td>
<td>-0.002</td>
<td>32</td>
</tr>
<tr>
<td>PRODOM PROFILE</td>
<td>-0.034</td>
<td>34</td>
</tr>
<tr>
<td>PRODOM PROSITE</td>
<td>0.159</td>
<td>14</td>
</tr>
<tr>
<td>PRODOM SMART</td>
<td>0.048</td>
<td>28</td>
</tr>
<tr>
<td>PRODOM TIGRFAMs</td>
<td>0.036</td>
<td>29</td>
</tr>
<tr>
<td>PROFILE PROSITE</td>
<td>0.070</td>
<td>25</td>
</tr>
<tr>
<td>PROFILE SMART</td>
<td>0.493</td>
<td>1</td>
</tr>
<tr>
<td>PROFILE TIGRFAMs</td>
<td>-0.201</td>
<td>36</td>
</tr>
<tr>
<td>PROSITE SMART</td>
<td>0.155</td>
<td>15</td>
</tr>
<tr>
<td>PROSITE TIGRFAMs</td>
<td>0.132</td>
<td>18</td>
</tr>
<tr>
<td>SMART TIGRFAMs</td>
<td>-0.189</td>
<td>35</td>
</tr>
</tbody>
</table>
Table 3 - 4 Decision Tree and Naïve Bayes orthologue prediction performances for five different datasets, including UniGene, Clutr and KOG.

The size-based domains with a domain size of 300 amino acids had consistently the best performance in terms of the average number of correct predictions based on 10-fold cross validations of each group of orthologues and Swiss-Prot sequences with matching and non-matching annotations (maximum 40 sequences, 20 matching, 20 mismatching – balanced data sets).

<table>
<thead>
<tr>
<th>Domain type</th>
<th>Domain size</th>
<th>Decision Trees</th>
<th>UniGene</th>
<th>Inparanoid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both</td>
<td>100</td>
<td>54.3%</td>
<td>76.3%</td>
<td>75.8%</td>
</tr>
<tr>
<td>Both</td>
<td>150</td>
<td>55.0%</td>
<td>76.3%</td>
<td>73.8%</td>
</tr>
<tr>
<td>Both</td>
<td>300</td>
<td>54.6%</td>
<td>75.2%</td>
<td>75.6%</td>
</tr>
<tr>
<td>InterPro</td>
<td>-</td>
<td>54.7%</td>
<td>80.2%</td>
<td>79.6%</td>
</tr>
<tr>
<td>Size</td>
<td>100</td>
<td>62.8%</td>
<td>82.3%</td>
<td>81.8%</td>
</tr>
<tr>
<td>Size</td>
<td>150</td>
<td>71.0%</td>
<td>81.7%</td>
<td>82.8%</td>
</tr>
<tr>
<td>Size</td>
<td>300</td>
<td>84.9%</td>
<td>82.6%</td>
<td>85.8%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td></td>
<td>75.8%</td>
<td>73.7%</td>
<td>74.9%</td>
</tr>
<tr>
<td>Total sequences</td>
<td></td>
<td>1,000</td>
<td>964</td>
<td>23,260</td>
</tr>
<tr>
<td>Orthologue groups</td>
<td></td>
<td>1,000</td>
<td>964</td>
<td>11,630</td>
</tr>
</tbody>
</table>
Table 3 - 5 Incorrect BLAST hits for 1,000 randomly selected human UniGene against Swiss-Prot and the mouse, dog and rat UniGene databases.

<table>
<thead>
<tr>
<th>No.</th>
<th>Gene Name</th>
<th>Predicted:</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Sperm associated antigen 17</td>
<td>Canis familiaris similar to projection</td>
</tr>
<tr>
<td>38</td>
<td>Rap guanine nucleotide-exchange factor (GEF) 1</td>
<td>Rattus norvegicus partial mRNA for C3G protein (c3)</td>
</tr>
<tr>
<td>86</td>
<td>Palmitoyl-protein thioesterase 2</td>
<td>Multiple EGF-like-domain protein 8 protein (Vasc)</td>
</tr>
<tr>
<td>143</td>
<td>Leukotriene C4 synthase</td>
<td>Mastemund-like protein 1</td>
</tr>
<tr>
<td>144</td>
<td>&quot;Neuroblastoma breakpoint family, member 1&quot;</td>
<td>Trichothyalin</td>
</tr>
<tr>
<td>177</td>
<td>&quot;Phosphatidylinositol glycan anchor biosynthesis&quot;</td>
<td>Endothelial zinc finger protein induced by tumor</td>
</tr>
<tr>
<td>179</td>
<td>Zinc finger protein 138</td>
<td>Canis familiaris similar to selenophos</td>
</tr>
<tr>
<td>189</td>
<td>Rheumatoid factor RF-ET12</td>
<td>Ig heavy chain V-III region HPC7s (Fragment)</td>
</tr>
<tr>
<td>218</td>
<td>Chemokine (C-C motif) receptor 10</td>
<td>Rattus norvegicus G protein-coupled rec</td>
</tr>
<tr>
<td>237</td>
<td>Transmembrane protein 106A</td>
<td>Canis familiaris similar to CCAAT displ</td>
</tr>
<tr>
<td>287</td>
<td>Exocyst complex component 7</td>
<td>Rattus norvegicus rex070 mRNA</td>
</tr>
<tr>
<td>300</td>
<td>&quot;Mannosidase, alpha, class 2B, member 1&quot;</td>
<td>Canis familiaris similar to mitogen-act</td>
</tr>
<tr>
<td>302</td>
<td>Nucleoporin 85kDa</td>
<td>Canis familiaris similar to pericentrin</td>
</tr>
<tr>
<td>310</td>
<td>MAP/microtubule affinity-regulating kinase 2</td>
<td>Rattus norvegicus mRNA for serine/threonine kinase MARK</td>
</tr>
<tr>
<td>329</td>
<td>Doppy family member 1</td>
<td>Canis familiaris similar to paf-1-like</td>
</tr>
<tr>
<td>331</td>
<td>Dual oxidase maturation factor 1</td>
<td>Canis familiaris similar to Namb-astera</td>
</tr>
<tr>
<td>347</td>
<td>Fibronectin leucine rich transmembrane protein</td>
<td>Rattus norvegicus Shal-related potassium channel K</td>
</tr>
<tr>
<td>374</td>
<td>Neuronal growth regulator 1</td>
<td>Rattus norvegicus mRNA for Klon</td>
</tr>
<tr>
<td>381</td>
<td>Tetraspanin 32</td>
<td>Phex protein</td>
</tr>
<tr>
<td>450</td>
<td>Neuronal PAS domain protein 4</td>
<td>Canis familiaris similar to HLH-PAS tra</td>
</tr>
<tr>
<td>458</td>
<td>Cathepsin C</td>
<td>LINE-1 reverse transcriptase homolog</td>
</tr>
<tr>
<td>464</td>
<td>&quot;Inositol polyphosphate-5-phosphatase, 40kDa&quot;</td>
<td>Rattus norvegicus Nk6 transcription fac</td>
</tr>
<tr>
<td>471</td>
<td>&quot;Phospholipase A2 receptor 1, 180kDa&quot;</td>
<td>Lymphocytic antigen 75 precursor (DEC-205) /CD205 a</td>
</tr>
<tr>
<td>499</td>
<td>C1q and tumor necrosis factor related protein9</td>
<td>Otolin-1 precursor</td>
</tr>
<tr>
<td>518</td>
<td>Free fatty acid receptor 3 (FFAR3)</td>
<td>Potative G-protein coupled receptor 41</td>
</tr>
<tr>
<td>552</td>
<td>solute carrier family 4</td>
<td>Anion exchange protein 4 (Anion exchanger 4) /Sodi</td>
</tr>
<tr>
<td>563</td>
<td>5-methyltetrahydrofolate-homocysteine methyltransferase</td>
<td>Rattus norvegicus similar to methionine</td>
</tr>
<tr>
<td>590</td>
<td>mRNA for Ras association (Raf1GUS/AF-6) domain family</td>
<td>Rattus norvegicus Mapx1 mRNA</td>
</tr>
<tr>
<td>601</td>
<td>Fc epsilon receptor III mRNA</td>
<td>C-type lectin superfamily member 6 (Dendritic cell)</td>
</tr>
<tr>
<td>612</td>
<td>vitram D receptor-interacting protein (DBR92)</td>
<td>Rattus norvegicus thyroid hormone recep</td>
</tr>
<tr>
<td>644</td>
<td>integrin</td>
<td>LINE-1 reverse transcriptase homolog</td>
</tr>
<tr>
<td>683</td>
<td>NADH dehydrogenase (ubiquinone) 1 alpha subcomplex</td>
<td>Rattus norvegicus mRNA for 14-3-3 protein gamma-su</td>
</tr>
<tr>
<td>713</td>
<td>tyrosine 3-monooxygenase/tryptophan 5-monooxygenase</td>
<td>Rattus norvegicus mRNA for 14-3-3 protein gamma-su</td>
</tr>
<tr>
<td>714</td>
<td>E2-k-alpha-ketoglutarate dehydrogenase complex</td>
<td>Rattus norvegicus dihydrofolate reductase</td>
</tr>
<tr>
<td>735</td>
<td>cytochrome P450</td>
<td>25-hydroxyvitamin D-1 alpha hydroxylase, monochon</td>
</tr>
<tr>
<td>743</td>
<td>mucin (MUC3) mRNA</td>
<td>Histidine-rich glycoprotein precursor</td>
</tr>
<tr>
<td>774</td>
<td>otoratin (OTOR)</td>
<td>Mus musculus mRNA for melanoma inhibitory activity</td>
</tr>
<tr>
<td>805</td>
<td>aldehyde dehydrogenase 18 family</td>
<td>Rattus norvegicus pyrroline-5-carboxyla</td>
</tr>
<tr>
<td>928</td>
<td>phosphoribosylaminomiazolase carboxylase</td>
<td>Mus musculus similar to Multifunctional</td>
</tr>
<tr>
<td>953</td>
<td>regulatory solute carrier protein</td>
<td>Rattus norvegicus pleckstrin homology d</td>
</tr>
<tr>
<td>961</td>
<td>protease</td>
<td>Prestasin precursor (EC 3.4.21.-)</td>
</tr>
<tr>
<td>977</td>
<td>receptor-interacting serine-threonine kinase 4</td>
<td>Canis familiaris similar to ankyrin rep</td>
</tr>
<tr>
<td>981</td>
<td>peroxisome proliferative activated receptor</td>
<td>Rattus norvegicus PPAR delta protein (PPAR delta)</td>
</tr>
</tbody>
</table>
Table 3 - 6 Swiss-Prot protein Decision Tree lookup and prediction performance.

Among 53,895 unique Swiss-Prot protein annotations, 82% (44,345 groups) traced back to itself based on a lookup of consecutive motif pairs compiled from up to 10 members of the functional group. Of the 65% of groups with tree predictions, 88% (57% of all groups) were predicted as matches. A total of 14% protein groups were predicted as belonging solely to the original functional group and 43% that were also predicted as belonging to at least one other functional group. A subsample of 500 bp fragments from the 5-prime end of the first sample had 93% successful lookups to the original functional group. The 500 bp fragments had correct prediction and unique correct prediction rates of 78% and 41% respectively.

<table>
<thead>
<tr>
<th></th>
<th>Swissprot</th>
<th>500 bp Fragments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree lookups</td>
<td>53,895</td>
<td>100%</td>
</tr>
<tr>
<td>Correct tree lookups</td>
<td>44,345</td>
<td>82%</td>
</tr>
<tr>
<td>Tree predictions</td>
<td>51,141</td>
<td>95%</td>
</tr>
<tr>
<td>Total self predictions</td>
<td>34,897</td>
<td>65%</td>
</tr>
<tr>
<td>'True' self predictions</td>
<td>30,831</td>
<td>57%</td>
</tr>
<tr>
<td>'False' self predictions</td>
<td>5,107</td>
<td>9%</td>
</tr>
<tr>
<td>Unique correct predictions</td>
<td>7,396</td>
<td>14%</td>
</tr>
<tr>
<td>Multiple predictions</td>
<td>39,576</td>
<td>43%</td>
</tr>
</tbody>
</table>

The table above summarizes the performance of the Decision Tree lookup and prediction for Swiss-Prot protein annotations and 500 bp fragments. The performance metrics include tree lookups, correct tree lookups, tree predictions, total self predictions, 'true' self predictions, 'false' self predictions, unique correct predictions, and multiple predictions.
Figure 3 - 1 Swiss-Prot training/test set selection based on InterPro hits of orthologue set.

The InterPro hits of all sequences in the orthologue set are used to extract Swiss-Prot sequences with similar combinations of motifs. Even numbers of ‘match’ and ‘mismatch’ sequences are collected, then aligned with ClustalW.
The multiple alignment of orthologue and Swiss-Prot sequences is used to position the orthologue sequences relative to each other. The superdomains are calculated based on the positions of InterPro hits on the orthologue sequences. Attribute values for all sequences are allocated as follows: 1) If sequence contains a motif belonging to the domain, the attribute value is the name of the domain. 2) If it contains a different motif, the attribute value is the concatenation of all motif names. 3) If not motif is present, the attribute value is "*".
Figure 3 - 3 Decision Tree generation using InterPro attributes.
Figure 3 - 4 A workflow for comparing Naïve Bayes and Decision Tree models of motif signatures in three orthologue databases and a 1,000-sequence sample of the human UniGene EST collection.

InterProScan is used to find the InterPro motifs at the ‘Orthologues Iprscan’ stage in the upper left portion of the workflow. For each orthologue group, Swiss-Prot sequences with the same function and different functions that share the most InterPro motifs in common with the orthologue group are added to the orthologues to form the Iprscan Set. Each Iprscan Set is then aligned with ClustalW and attribute domains are constructed based on boundaries of InterPro hits in the multiple alignment and/or a fixed length. The machine learning attribute values correspond to the longest InterPro motif name within each domain (or ‘*’ if no InterPro motifs are present). The sequence classification is ‘1’ for sequences with the same function as the orthologue group and ‘0’ for sequences with different functions. For each orthologue group, 10-fold cross validation was carried out using balanced sets of 20 ‘1’ and 20 ‘0’-classified sequences. The same training and test sets were used for the Naïve Bayes and Decision Trees classifiers. Finally, summary statistics of the average prediction accuracies from the 10-fold cross validations are calculated for both Naïve Bayes and Decision Trees.
Figure 3 - 5 Most InterPro motifs occur in two or more functionally different Swiss-Prot proteins.

Graph A shows that, of the 24,988 unique InterPro motifs found in the Swiss-Prot collection, 10.8% (2,689 motifs) appear in only one protein. For the remainder, 27.3% (6,833 motifs) are found in two to five different proteins, 41.4% (10,341 motifs) are found in six to fifty different proteins and 20.5% (5,125 motifs) are found in more than 50 different proteins. The breakdown by unique functional group (containing one or more proteins) is shown in graph B, showing that 29.1% (7,322) of InterPro motifs belong to a single Swiss-Prot functional group.
Chapter 4: Improved EST database annotation
Abstract

Expressed sequence tags (EST), partial sequences of cDNA (mRNA converted to DNA), provide the most compelling evidence for the existence and activity of genes. For many non-model species, EST databases are the first genomic resource to be developed and EST deposits make up the majority of sequences in GenBank. Automated and accurate annotation is a first step to realizing the potential of EST resources. However, current EST pipelines lack quality control mechanisms and have limited usability. The user’s ability to interact with the data is hindered by inadequate or absent web interaction tools and a lack of scientific workflow support. To leverage the benefits of inexpensive computing clusters, EST pipelines must also support distributed computing in a robust and fault tolerant manner.

To meet the above demands, we have developed an updated version of the Fundulus heteroclitus EST database (Funnybase), a freely-accessible online resource for marine genomics. Release 2.0 of Funnybase was produced using the myEST platform, a comprehensive and integrated EST quality control, assembly and analysis tool designed to complement and extend existing EST analysis tools and technologies.
Introduction

Expressed sequence tags (ESTs) in the public dbEST database comprise approximately 61% (51 million sequences in March 2008) of the total 83 million sequences in GenBank, a rapid increase from the initial dbEST collection of 10,000 ESTs (10% of GenBank) in 1992 (Boguski, et al. 1993, NCBI 2007, GenBank 2008). The growth in ESTs has outstripped that of the non-EST sequences within GenBank, which has doubled in size every 18 months from 1982 to the present. ESTs provide an important genomic resource for comparative and functional genomics studies, identifying genes in newly sequenced genomes, tissue-specific expression analysis, single nucleotide polymorphism (SNP) prediction and metabolic pathway discovery. There are currently 825 organism EST collections with more than 1,000 sequences, the vast majority of which are so-called non-model organisms whose genomes have not been sequenced (Figure 4 - 1).

While there is a growing use of EST databases for non-model organisms, there are few resources available for integrating EST assembly, visualization, annotation and analysis. And, in particular, there is a paucity of open-source tools for improving the quality of EST databases during their construction and annotation, particularly for the small to medium-sized EST databases which represent the majority of organisms in dbEST. This paper describes the use of quality control methods within an integrated EST assembly and web interface platform to improve the quality of EST collection sequences and their annotation, and to provide improved user access to and interaction with these data.
2.1 Quality control is essential for EST projects

Current EST sequencing, assembly and annotation technologies are prone to certain errors, which can adversely affect the quality of the resulting EST collections that are directly submitted to dbEST by laboratories world-wide and are not curated. Aside from errors in sequencing, which can be partially mitigated by filtering based on quality values, traditional EST clustering, or assembly, methods often give incorrect results because they fail to consider alternative splicing (Heber, et al. 2002). High error rates in the assembly of EST collections may be a source of the approximately 40% of human ESTs that do not match identified genes (Larsson, et al. 2005) although a small portion of this may be explained by non-gene transcription units (TUs); non-coding RNAs implicated in mRNA processing, transcription factor recruitment and chromatin remodeling (Lipovich and King 2006).

Previously, it was found that 26% of pairs of human ESTs from dbEST did not match human genomic sequences (Wolfsberg and Landsman 1997) and artifacts in EST generation were suspected to be the cause of some of these unusual ESTs although. Compounding any possible error due to incorrectly sequenced or assembled ESTs, BLAST-based annotations can yield dubious annotations if true orthologues are missing from BLAST target databases even at commonly used BLAST cutoff thresholds (Young and Crawford 2008). This issue is particularly important due to the risk that errors in EST sequence annotations can be propagate or cascade through other databases given the overwhelming reliance of database annotation on computational sequence comparison.

There is a need to reduce the risk of inaccurate annotations due these factors: poor quality
input sequences, inappropriate assembly and inaccurate annotation. To this end, the myEST assembly and annotation pipeline was designed to provide quality control measures before, during and after EST assembly, and at the subsequent annotation phase of the completed collection, and was implemented using EST collections of three very different marine organisms – *Fundulus heteroclitus*, *Karenia brevis* and *Aplysia californica* (Figure 4 - 2).

### 2.2 Improved data access and interaction

Another important facet of EST database usage addressed by the myEST platform is access to and interaction with the raw and processed data and parameterization during the various steps in the production pipeline. Many large EST resources, such as dbEST, TIGR and the Joint Genome Institute (JGI), currently provide only basic statistics on EST assembly procedures and results, and limited information on sequence annotation procedures (Grid Middleware Development Group 2007). At the other end of the spectrum, EST assembly platforms commonly provide a variety of functionalities (Yee and Conklin 1998, Chevreux, et al. 2004, Kumar, et al. 2004, Lazzari, et al. 2005, Xu, et al. 2003) but, with few exceptions (Lee, et al. 2007), have limited or no sequence annotation nor annotation quality control capabilities. Almost all EST resources rely on relational databases but lack simple-to-use interfaces to access all of the data.

Ideally, a database interface should help users with limited programming experience to explore data, visualize it and interpret the results. Interactive workflow systems in the biological sciences have the potential to enable more efficient use of bioinformatics tools and
free up time and energy for more creative scientific activities. In order to achieve this, workflow systems should provide most or all of the following:

- Automation of repetitive tasks
- Data integration and visualization
- Enabling of data sharing
- Extensible/customizable and simple graphical user interface
- Allow user to generate, run, track and save workflows
- Track and associate results with workflow stages
- Fault tolerance, particularly with distributed computing
- Comparison of workflows
- Reusable generic features
- Extensibility for the expert user
- Publication of results

### 2.3 Data sharing and interaction across the web

Given the need for visualization and data sharing, the ability to interact with data over the Internet is an essential feature of a workflow system. AJAX (Asynchronous Javascript and XML) is an important tool for creating web applications that have the both the advantages of web pages (open access, communication and sharing of information) and desktop applications (complex and highly interactive user interface). Highly interactive web applications are made possible by updating parts of a web page without the need to reload the page. An Ajax engine, such as a Javascript AJAX library, acts as an intermediary passing requests and responses between client and server, and updating portions of the web page
without reloading the whole page (Figure 4 - 3), hence the term ‘asynchronous’. The growing popularity of AJAX among web designers is due to its simplicity relative to other techniques such as Simple Object Access Protocol (SOAP) or Remote Procedure Call (RPC). The technologies used by AJAX are mature, stable and ubiquitous - they are supported by all major web browsers and there are many open source libraries of standardized, reusable code which are freely available.

**Methods**

### 3.1 Organism EST collections

The *F. heteroclitus* EST collection is comprised of contigs and singlets assembled from two groups of reads: 1) Assembled contigs and singlets from 34,810 newly sequenced EST reads from a range of tissue sources (Table 4 - 1), and 2) Contigs (342 EST clusters) and singlets (451 ESTs) from Funnybase 1.0 (Paschall, et al. 2004) with BLAST annotations not found in (1) above. The Funnybase 1.0 assembly was based on 66,554 reads (44,372 with phred quality values and 22,182 without phred values).

The *K. brevis* and *A. californica* collections were assembled using 22,351 and 6,901 reads with quality values, respectively. CAP3 was used for clustering all collections of reads, with or without accompanying phred quality scores. Uni-directional cloning was used for all three species: the vector pSmart-Lucigen was used for *K. brevis* and *A. californica*, and vectors pSmart-Lucigen, pBluescriptIISK, pSmart-HCKan and bluescript were used for *F. heteroclitus*. 
3.2 BLAST and InterPro motif searches

All BLAST and InterPro motif searches were carried out on a 32-CPU Apple Xserve G5 cluster (dual 2.3 GHz CPU, L2 Cache 512 KB, 1 GB memory and 1.15 GHz Bus Speed). BLASTP analyses were performed for all three EST databases against the Swiss-Prot protein database and the human, mouse, rat and dog RefSeq protein databases using the default parameters and filtering with seg. The Perl5 InterproScan application was used to scan for InterPro motifs for all databases (100 hours to scan approximately 12,000 sequences). All of the public InterPro databases were used in our analysis except for SUPERFAMILY which was excluded because it was computationally most costly (by a factor of ten) than most of the other InterPro motif database searches and due to a significant correlation of occurrence of SUPERFAMILY motifs and motifs of other databases in the EST sequences.

Based on InterProScan searches using the *F. heteroclitus* EST collection, it was found that InterProScan run durations were not strongly correlated with the minimum translated length threshold setting within the range 50 to 80 amino acids. Since the increased stringency of the 80 amino acid cutoff did not result in a significant time saving, the more permissive 50 amino acid threshold was used. The two command line parameters ‘-iprlookup’ and ‘-goterms’ for the lookup of InterPro IDs and Gene Ontology (GO) terms, respectively, were also used because they were required for our analysis, although the InterProScan runs were significantly slower as a result.

Since cloning was unidirectional, the ‘-noreverse’ flag was also set within InterProScan for
the ‘sixpack’ ORF translation application in order to limit output to the three reading frames on the sense strand. Default settings were used for all other InterProScan parameters (see information on using InterProScan: http://www.ebi.ac.uk/interpro/user_manual.html).

InterProScan jobs were allocated and monitored using the Sun Grid Engine (SGE) job scheduler and additional Perl5 software tools with the myEST package were used to mitigate issues such as job failure and node failure, while maximizing the efficiency of cluster usage.

**Results**

The EST collections of the three marine species *K. brevis*, *A. californica* and *F. heteroclitus* with read libraries of size 6,901, 22,351 and 68,094 reads (including 29,326 new reads), were assembled and analyzed within from 24 hours to 5 days using an EST assembly pipeline comprised of a series of workflows for read processing, assembly and annotation and with multiple quality control measures carried out at each of the following three steps in the myEST pipeline: pre-assembly, assembly and post-assembly (Table 4 - 2).

**4.1 Pre-assembly quality control**

Pre-assembly plate checking was the first quality control step to avoid incorrect read ID designations due to accidentally rotated or swapped plates in EST collections. This is an essential feature for large collections that lack barcoding or other plate identification measures. Prior to EST assembly, all sequences in each plate are compared against the marker sequences by BLASTN search. According to the predetermined static and dynamic
(plate-dependent) marker positions, plates may be diagnosed as rotated, swapped or both. Read sequences can be downloaded by clicking on any non-empty well in the graphical representation of the 96-well plate. The plate viewer function of the web interface displays the well contents of each 96-well plate in the selected sequence collection (Figure 4 - 4). In order to ascertain the cost-benefit of sequencing additional sequences - whether or not more ESTs will add significantly more unique clusters to the database - a cumulative assembly using *A. californica* EST reads was carried out to assess the level of ‘saturation’ of the EST library (Figure 4 - 5). Five equal portions of the total of 6,091 reads were added sequentially to the database, which was reassembled each time to determine if the number of new, unique sequences increased significantly as more read sequences were added to database. Each successive assembly was carried out using CAP3 (default parameters). It was found that the number of assembled sequences continued to rise relatively steadily, as was expected with a library of this size (Figure 4 - 5). The number of assembled sequences that had BLAST hits against Swiss-Prot and the Refseq databases (human, mouse, dog and rat) also rose relatively linearly with the increase in EST reads (Table 4 - 3), supporting the conclusion that the ‘elbow’ or flattening off indicating maximal utility had not been reached for this EST database.

In an additional step to increase the stringency of sequence screening, sequence read filtering was also performed after vector sequence screening by masking out vector sequence with ‘X’s. The workflow excises all residues downstream of A) a second ‘poly-X’ sequence
appearing in the sequence indicating the presence of additional vector sequence, and B) any polynucleotide sequence of more than 10 residues.

4.2 Quality control during cluster assembly

Optimization of CAP3 assembly options was also used to assess the balance between avoiding incorrect clustering and maximizing the number of correct clusters. Overly permissive clustering thresholds can result in anomalous clusters of unrelated sequences, which may in turn result in conflicting annotations for the clusters. The effect of changes in three CAP3 parameters was investigated (Table 4 - 4) and two of these were found to significantly affect the numbers of contigs and singlets produced: the overlap percent identity cutoff (-p) and the overlap length cutoff (-o). These two parameters were varied together in a Latin square design to assess their combined effects (Table 4 - 5). On manual inspection of the contig alignments, it was found that lower values of the overlap percent identity cutoff produced significant numbers of dubious contigs, as defined by clusters in which two or more groups of sequences were barely overlapping and showed no overall sequence similarity. As a result, the overlap percent identity and overlap length cutoffs were set to parameters were set to 90% and 100 bp, respectively.

4.3 Post-assembly quality control

Quality-value filtering of contigs and singlets before inclusion in the EST collection was also assessed. Before the assembled contigs or singlets were entered into the EST collection, all sequences with average quality values below the threshold were discarded. Overall, pre-
assembly filtering of polynucleotide sequences improved the number of BLAST hits in the completed collection had a much greater effect on the number of BLAST hits than CAP3 parameterization or post-assembly quality-based filtering (Table 4-6). As a result of these findings, pre-assembly filtering was used for the assemblies of all three organisms (F. heteroclitus, K. brevis and A. californica), and CAP3 parameters were set as follows: overlap identity cutoff (-p) = 90%, overlap length (-o) = 100bp. A post-assembly length filter of 50bp and a quality filter cutoff of 20 for contigs and 15 for singlets was also used for all assemblies, with the exception of F. heteroclitus where a contig quality cutoff of 10 was used.

Manual inspection of assembly statistics was also an important part of the quality control process. For each EST collection database, detailed statistics of the completed assembly are automatically calculated and displayed in the web interface (Figure 4-6). The quality values of the assembled contigs and singlets can also be viewed as a count of sequences meeting a user-defined quality threshold or as a bar graph distribution of contig and singlet phred quality values (Figure 4-7). This indicator of sequence quality can be used to set quality cutoffs for extracting sequences as microarray probes. Another factor relevant to selecting microarray probes is the number and distribution of sequences supporting a contig consensus sequence. Two metrics of coverage - read depth and read coverage – were devised and calculated for all contigs in each collection (Figure 4-8). These values can be also used in the selection of high-coverage reads for EST library subtractions or normalizations.
The 3-frame translation of all collection sequences is displayed on the Search page of the myEST interface. This was sufficient for the purposes of annotating the three species databases; the BLAST and InterProScan algorithms employed in the pipeline do not require peptide sequence input. Should more accurate ORF prediction be required, for example for further analysis of unannotated sequences, there are many ORF prediction applications (Parkinson and Blaxter 2004, Fukunishi and Hayashizaki 2001) designed to cope with the false base calls and frameshifts that are a common artefact of the EST sequencing process. Any of these applications could be easily incorporated into the platform.

4.4 Choice of BLAST database

The BLAST-based high-throughput annotation of sequence databases can be improved by comparing BLAST hits from two databases, such as Swiss-Prot and UniGene, with as much as 99% confidence for annotations where the best BLAST hits from both databases are in accord (Young and Crawford In preparation). Swiss-Prot and RefSeq were used for annotation because of their relative abundance of annotations containing descriptions of protein function. The BLAST annotation portion of the myEST pipeline begins with a BLASTX sequence comparison against Swiss-Prot and four Refseq protein databases (human, dog, mouse, rat). This choice of databases was based on a manual review of the annotations of BLAST hits for the complete F. heteroclitus EST collection against the following databases: GenBank non-redundant (nr), UniGene (human, mouse, rat, dog), Ensembl (C. elegans, chimpanzee, fugu, human, mouse, tetraodon, yeast), Swiss-Prot and Refseq (human, mouse, rat, dog). Because of its higher number of annotations containing
gene function information, Swiss-Prot was considered the first tier database and the Refseq mammalian databases as the second tier databases. The Refseq mammalian sequences (human, mouse, dog, rat) were chosen to minimize the risk of varying evolutionary distance complicating the comparison between E-values of hits in the different Refseq databases. The assumption was that these four mammalian genomes are equidistant from *F. heteroclitus* and other non-mammal marine organisms in terms of evolutionary divergence. In addition to the top two database tiers, a third tier of databases containing sequences from dinoflagellates and other evolutionarily similar species was used for the annotation of *K. brevis* sequences.

Another reason for preferring Swiss-Prot over other databases is ease of automated processing of annotations. Swiss-Prot follows a strictly adhered to annotation format, whereas many other databases have much less structured annotation formats that are not strongly enforced. For example, simple parsing rules are insufficient to distinguish between synonyms in Refseq annotation (Table 4 - 7). Furthermore, annotations in Swiss-Prot generally refer directly to biological function, whereas other databases often include less useful information.

The annotation procedure was as follows:

1. Obtain BLAST hits against Swiss-Prot and Refseq databases with an E-value cutoff of $10^{-5}$

2. For each collection sequence:

   If both Swiss-Prot and Refseq hits available:
Use Swiss-Prot annotation if

\[ \ln (\text{Swiss-Prot E-value}) < 1.2 \times \ln (\text{Refseq E-value}) \]

Otherwise, use Refseq annotation.

Otherwise, use the annotation from the lowest E-value hit

3. Run InterProScan protein motif search on all sequences

4. For sequences with no BLAST hits, use InterProScan hit annotation if available

5. Calculate support values for annotation based on matching annotation from the three different databases (e.g., two out of three matching annotations gives a support value of 2 out of a possible 3).

6. Calculate ‘unique hits’ – the total number of unique best BLAST hits (some collection sequences may have been allocated their annotation from the same best BLAST hit sequence.

For both *F. heteroclitus* and *K. brevis*, the number of ‘unique hits’ – BLAST hits against a unique target sequence - was much less than the total number of collection sequences with BLAST hits. *F. heteroclitus* has 2845 unique sequences versus 5184 annotated sequences, and *K. brevis* has 759 unique versus 1024 annotated sequences. Annotated sequences that shared the same best BLAST hits also had high average
quality values, even in regions of sequence divergence, were considered possible SNPs or gene duplications and so were retained as separate entities in the collection.

4.5 Incremental mismatch probability aids Annotation

In addition to comparing BLAST hits in high quality databases, the myEST pipeline also employs an incremental mismatch probability methodology based on BLAST hit distributions for prospective orthologues which has been shown to detect approximately half of incorrect annotations by providing a statistical reference for assessing the significance of BLAST hits (Young and Crawford In preparation). The incremental mismatch probability returns the likelihood that the best BLAST hit at the particular E-value for a collection sequence occurred by chance, i.e., is a false positive. Whereas an E-value is the expected number of random matches of a certain score within a database of a particular size, the incremental mismatch probability is the likelihood that two proteins matched with a particular E-value are in fact functionally similar.

One of the drawbacks of the incremental mismatch probability is that it relies upon the completeness and integrity of the annotations found within the target database; ideally, all sequences must be annotated and annotations must be consistent (only one annotation for the same function) but not necessarily non-redundant (one protein may have several functions and therefore annotations). This can be assumed to a certain extent with the Swiss-Prot database because it is a multiple-species database of manual annotations based on a controlled vocabulary, and to a lesser extent with the RefSeq database. However, the
incremental mismatch probability method would be less suitable for use with databases that are less stringent in terms of annotation format and evidence standards, or that contain a significant proportion of functionally non-descriptive annotations such as clone IDs. For example, UniGene mammalian databases have high rates of non-descriptive annotation (Young and Crawford *In preparation*) and there is a very significant heterogeneity in the consistency, accuracy and format of annotations in the GenBank non-redundant (*nr*) sequence database, which consists of sequences uploaded by thousands of different sequence donors with little or no quality control or curation.

### 4.6 Mappings between external databases

The EST collection pipeline makes extensive use of multiple mapping tables for external database two purposes: 1) Extracting metabolic pathway annotation, and 2) Finding annotation synonyms. Mapping involves linking two different classifications systems, usually in the form of lists of paired unique sequence identifiers, one from each classification systems. The available external database mappings of the Gene Ontology consortium (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used to produced the mappings (Table 4 - 8).

#### 4.6.1 Metabolic pathway annotation

Of the 4,242 Swiss-Prot BLAST hits (representing 3,313 unique Swiss-Prot sequences) that were used to accord putative annotations for *F. heteroclitus* EST collection sequences, only 1,424 contained Enzyme Commission (EC) IDs in their annotation. In order to increase the
A number of EC identifiers and derive GO and KEGG identifiers for the collection sequences, an extensible mapping system was developed. For an unlimited number of input mapping files, the application searches the complete space of all possible non-redundant ‘mapping chains’ between any source database and target database. For example, a possible mapping chain between EC numbers and Pfam IDs could be EC $\rightarrow$ GO $\rightarrow$ Pfam. For each input EC IDs, the application looks up the ‘ec2go’ mapping to find a corresponding GO ID. If a GO ID is found, the ‘go2pfam’ mapping would be search to find at least one Pfam ID corresponding to the GO ID. Note that, depending on the source and target databases, this could result in a One:Many, Many:One or Many:Many relationship. The maximum chain length can be specified by the user or the default value is used (3 databases, i.e., 2 mapping steps).

An exhaustive search of all combinations of available mappings from Swiss-Prot and Refseq BLAST hit IDs to GO, KEGG and EC IDs yielded 2,361 Enzyme Commission numbers, 114,266 GO IDs and 41,569 KEGG IDs for the *F. heteroclitus* EST collection. This is available as a downloadable file and will later be incorporated into the Search display for each sequence. To aid the interpretation of these mappings, all intermediate steps in each chain of mappings will be included in the display.

In order to find more specific gene function GO terms that appear further down the tree of GO annotations, the mapping-derived GO IDs were filtered by removing all GO entries that are also found in the GO Slim ontology, a reduced list of top-level, general GO terms. As a
footnote, a new International Protein Identifier (IPI) mapping with GO is being developed to provide a unifying index to major databases covering the proteomes of higher eukaryotic organisms, such as UniProt, Vega, Ensembl, RefSeq and TAIR. In the future, this could be used to replace or supplement the above GO-to-external mapping table. But IPI is still in its early stages; for example, there are currently only approximately 4,000 IPIs for human Refseq sequences.

4.6.2 Annotation synonyms

Synonyms for genes were found by searching for fuzzy matches in mapping tables, including Gene Ontology (GO) to external database mapping files. For example, a search for synonyms of ‘Putative 2Fe-2S ferredoxin’ yielded multiple synonyms with varying relevance values based on a fuzzy database match (Table 4 - 9). Although fuzzy matching of synonyms augmented by mapping information helps increase the list of search terms for matching annotations, there are two major constraints: 1) a lack of a consensus on annotation format hinders parsing into mapping entries and 2) for any particular target database, there may not be sufficient good-quality mapping information for the steps between the source (e.g., Swiss-Prot IDs, RefSeq IDs of BLAST hits) and target databases (e.g., Reaction IDs). Furthermore, even with the best quality information, fuzzy matching techniques have non-trivial false positive (Type I) and false negative (Type II) error rates. It is therefore not prudent to place too much emphasis on any particular mappings but, rather, to use them as a guide to indicate a general rate of similarity. For example, mappings can be used to improve the match and mismatch counts by reducing false negatives, in the calculation of the BLAST
distributions of the best BLAST hits of collection sequences.

4.7 Web Interface

The web interface was implemented in DHTML with client-side Javascript applications updating page elements in real time by AJAX-based interactions with Perl5 CGI applications on the server side. The interface is based on the classic design pattern for graphical user interfaces of all kinds: Model-View-Controller (Figure 4 - 9). The Model-View-Controller design pattern is comprised of three components: 1) a Model component located on the server side, in the form of the relational databases that comprise the EST sequence collections and the Perl CGI applications that interpret that model to reply to data requests from the web pages, 2) a View component consisting of combination of cascading style sheets (CSS) and HTML templates provides the View component, i.e., the layout and visual style of the website, and 3) a Controller component consisting of object-oriented Javascript modules that manipulate document object model (DOM) to achieve the functionality of an application in the web page. The Model-View-Controller pattern was implemented as a web application rather than a local desktop application in order to meet the requirement for sharing and communication of data, which was best achieved across the Internet.

Asynchronous Javascript and XML (AJAX) techniques are used to load an HTML template and its corresponding Javascript controller each time a new function is selected by the user (Figure 4 - 9). The interface provides most or all of the required functionality for small to medium-scale EST projects, where the emphasis is on flexibility and a minimal learning
curve. For users with more advanced programming abilities, myEST also provides solutions to issues of fault tolerance in distributed computing environments, workflow control and interactive web interface design that can be customized or incorporated into other systems.

The myEST platform comprises of a pipeline of workflows specifically tailored for EST database assembly and annotation. Each workflow accomplishes a key task in the EST analysis process, from reading sequence chromatograph files to fine-tuning sequence annotation results. Any selected workflow can be run and its progress monitored using the Workflow function (Figure 4 - 10). The user can also view all workflows in the database, or the component stages of the workflow, or view and edit the arguments for any application in any stage.

Although intended primarily for EST genomic pipeline analysis, the modular, object-oriented design of myEST allows it to be easily adapted or extended to meet the data display and analysis demands of any biological sequence pipeline. The application-style interface is intuitive for users familiar with desktop applications. Dropdown menus provide navigation through its various functions. Through the interface, users can create new databases of different types: EST, orthologue or custom sequence pipelines. For EST and orthologue databases, specialized workflows are initially added to the new database by default. The user can then add, modify, delete and start/stop workflows and monitor workflow progress.

4.8 Data visualization and exploration

In addition to the automated quality control checks described earlier, human exploration of
the data can detect errors in assembly or annotations, as well as biologically important aspects of the data. The Search function of the web interface provides a rich visual display of collection sequence information and related annotation data. Each collection sequence can be retrieved by its ID number, annotation, or by using external database IDs corresponding to its best BLAST hits. The advanced Search function performs searches using Boolean logic whereas the simple Search function uses fuzzy matching to retrieve inexact matches by order of relevance. Each retrieved sequence profile includes the sequence displayed with dark/light coding of quality values at each individual residue, predicted open reading frames (ORFs) and, if available, annotation information, and sequence displays of best BLAST hits from Swiss-Prot, Refseq (human, mouse, dog and rat), and the 12 InterPro motif databases. Protein domain information from the best Swiss-Prot BLAST hit is also displayed if available. For most sequence display elements, placing the mouse over the element will trigger a Javascript display of detailed information regarding the element. All of the BLAST and InterPro hit displays are linked to their external Internet source, which can be accessed by simply clicking on the sequence element.

The nucleotide sequences for an individual collection accession can also be downloaded directly from the sequence profile. By clicking another download arrow next to the scrolling list of search hits, the user can download all collection sequences that meet the particular search criterion. For EST sequences, the alignment of reads against the contig consensus sequence can also be viewed and downloaded.
The database can also be interrogated using protein or nucleotide sequences with the BLAST Search function. All collection database sequences, Swiss-Prot and the Refseq human, mouse, dog and rat databases can be searched in this manner. The search target can include a whole database or only specific sequences from different target databases; selected target sequences can be consecutively or simultaneously loaded into the target sequence box using their accession IDs and BLAST output results are ranked by E-value and displayed in a popup page.

Aside from the individual and multiple downloads available using the Search function, the sequence information of all EST reads and their related collection sequences can be downloaded as spreadsheet-compatible tab-separated values file using the Download function. Where available, BLAST hit information (hit ID and annotation, query and target start/stop, E-value, etc.) is also included for each sequence. The contents of the collection directory and its subdirectories can be viewed and any file can be downloaded. Users can also upload sequence files and BLAST databases using the Upload function.

### 4.7 dbEST Submission

Submission of annotated and unannotated ESTs to dbEST is an important and often required step for publishing the contents of a collection. Although a Perl sequence submission application is provided by dbEST, the preparation of dbEST submissions can be a non-trivial task. Instead, the user can input sequencing experiment information into myEST’s dbEST submission page which allows the user to input information on the sequencing experiment,
authors, publication, contact information, etc., and download the results in the correct format for dbEST submission. The input information is permanently saved and can be later retrieved and updated in the dbEST submission page.

**Conclusion**

We have described the EST assembly workflow as a series of quality control tasks designed to improve the quality of EST collection sequences and sequence annotations. Pre-assembly filtering of low-complexity regions in reads had a significantly increased the quality and number of BLAST annotations in the EST collections and the level of complexity of the assembled sequences. Assembly algorithm parameter adjustments and quality filters also had significant effects on sequence quality and the number of BLAST hits but at the expense of the overall quality of BLAST hits. Increased confidence in the quality of sequence annotation was achieved by the additional use of multiple BLAST target databases and an incremental mismatch probability approach for assessing the significance of BLAST hit E-values. Mappings between different external databases can be used to expand on the information gained from BLAST and InterPro protein-motif based annotations. Additional features include quality value distributions, sequence similarity comparisons between different databases and content-rich sequence searches.

The improved annotation and accessibility of Funnybase 2.0 and the *K. brevis* and *A. californica* EST collections increases their value as marine genomics resources. This was achieved by implementing a uniform set of quality control methods for these three
evolutionarily distant marine organisms which were sequenced with very similar technologies. However, different sequencing technologies or organisms may well show different levels and types of error, which underscores the importance of having a flexible and customizable framework. The myEST platform contains simple and flexible solutions for sequence databases issues such as workflow transparency, quality control, data access and interactive analysis, and will be made freely available as an open-source tool that may be useful for a wide range of sequence analysis projects.

The AJAX tools implemented in the web interface meet the scalability demands of a medium-size EST database by providing complex functions and visual displays with negligible processing time on the client browser. In particular, functions such as URL bookmarks are required to be highly scalable in terms of the volumes of data that can be stored because of the expected increasing complexity of visual presentations of workflows and sequence annotation information. Older or lower-specification computers may not be able to process the client-side Javascript fast enough to achieve the intended effect of smooth and seamless AJAX state transitions. However, Moore’s Law (CPU and memory chip capacity doubles every 18 months) ensures that processing power is a temporary concern. Low bandwidth may also impede the use of bioinformatics AJAX web applications but this problem would be more pronounced with traditional, page reloading websites. Both of the above points are relevant concerns with regard to users in developing countries, where computer hardware specifications may be lower and Internet connectivity speeds slower.
The open-source myEST platform can be applied to any EST project based on the most commonly used sequencing and EST assembly tools and it provides most or all of the required functionality for small to medium-scale EST collections, which make up the majority of deposits in the public dbEST database. Because of its modular design and SQL database-centered workflow architecture, the myEST platform can be easily adapted to meet different technology use and error mitigation strategies. For users with more advanced programming abilities, myEST also provides solutions to issues of fault tolerance in distributed computing environments, workflow control and an interactive interface design that can be customized or incorporated into other systems.
References


Table 4 - 1 Funnybase 2.0 EST tissue sample experiments and read counts.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Number Reads</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>126</td>
<td>5947</td>
<td>Heart and Liver cDNA from single Massachusetts’s population</td>
</tr>
<tr>
<td>130</td>
<td>8995</td>
<td>Liver cDNA from a population exposed to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>151</td>
<td>959</td>
<td>Heart cDNA from nine populations exposed to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>152</td>
<td>959</td>
<td>Liver cDNAs from nine populations exposed to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>153</td>
<td>1152</td>
<td>All developmental stages (whole animal)</td>
</tr>
<tr>
<td>154</td>
<td>768</td>
<td>Brain, bone and eye from nine populations exposed to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>155</td>
<td>942</td>
<td>Internal organs (spleen, gall bladder, kidneyS, fat) from nine populations exposed to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>156</td>
<td>954</td>
<td>Skin, fin and muscle from nine populations exposed to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>157</td>
<td>945</td>
<td>G-I tract and fatty tissue cDNAs from nine populations expose to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>158</td>
<td>956</td>
<td>Gill cDNAs from nine populations expose to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>159</td>
<td>12233</td>
<td>Mixed developmental, tissues and stressors from nine populations: all developmental stages (whole animal), brains, liver, heart, muscles, spleen, eye, GI, fat, gill, bone, gall bladder and epithelia expose to hypoxia, cold, heat, low and high salinity and hydrocarbons</td>
</tr>
<tr>
<td>Total</td>
<td>34810</td>
<td></td>
</tr>
</tbody>
</table>
Table 4 - 2 The three stages of quality control in the myEST pipeline.
The pipeline consists of a Pre-Assembly, Assembly and Post-Assembly stages which involve different types of quality control and error checking measures.

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-Assembly</strong></td>
<td></td>
</tr>
<tr>
<td>Plate checking</td>
<td>Verification of 96-well plate number and orientation</td>
</tr>
<tr>
<td>Read filtering</td>
<td>Removal of poor quality reads and low quality regions in reads</td>
</tr>
<tr>
<td>Cumulative assembly</td>
<td>Determine whether existing collection is ‘saturated’</td>
</tr>
<tr>
<td><strong>Assembly</strong></td>
<td></td>
</tr>
<tr>
<td>Parameter optimization</td>
<td>Assembly parameter optimization – maximize correct clusters and reduce incorrect clusters</td>
</tr>
<tr>
<td>Sequence filtering</td>
<td>Removal of poor quality or short contigs or singlets</td>
</tr>
<tr>
<td><strong>Post-Assembly</strong></td>
<td></td>
</tr>
<tr>
<td>Assembly statistics</td>
<td>Display detailed statistics of the completed assembly</td>
</tr>
<tr>
<td>Quality distributions</td>
<td>Display phred quality value distribution by sequence type (singlet vs contig)</td>
</tr>
<tr>
<td>ORF prediction</td>
<td>Display predicted open reading frames</td>
</tr>
<tr>
<td>Read coverage</td>
<td>Calculate density of read coverage for contigs</td>
</tr>
<tr>
<td>BLAST databases</td>
<td>Use a hierarchy of two or more target databases for BLAST annotation</td>
</tr>
<tr>
<td>Data visualization</td>
<td>Display assembled sequences with phred quality values, downloadable sequences and alignments</td>
</tr>
<tr>
<td>Data accessibility</td>
<td>Download whole-collection assembled sequences with functional and metabolic pathway annotation</td>
</tr>
</tbody>
</table>
Table 4 - 3 Expressed sequence tag (EST) reads from *A. californica* were successively and cumulatively assembled using CAP3. The number of contigs and singlets increased with the number of experiments making up the pool of pre-assembly EST reads, as did the number of assembled sequences (contigs and singlets) that had BLAST hits against Swiss-Prot and the RefSeq human, mouse, dog and rat databases (BLAST E-value cutoff: $10^{-5}$).

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Contigs</th>
<th>Singlets</th>
<th>Annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-20</td>
<td>22</td>
<td>876</td>
<td>261</td>
</tr>
<tr>
<td>1-30</td>
<td>25</td>
<td>1481</td>
<td>464</td>
</tr>
<tr>
<td>1-63</td>
<td>89</td>
<td>2391</td>
<td>760</td>
</tr>
<tr>
<td>1-79</td>
<td>106</td>
<td>3136</td>
<td>1046</td>
</tr>
<tr>
<td>1-89</td>
<td>115</td>
<td>3639</td>
<td>1150</td>
</tr>
</tbody>
</table>
Table 4 - 4 CAP3 cutoff parameters used in the optimization of *K. brevis* EST assembly.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range of values</th>
<th>Cutoff type</th>
</tr>
</thead>
<tbody>
<tr>
<td>-p</td>
<td>66-100</td>
<td>overlap percent identity</td>
</tr>
<tr>
<td>-s</td>
<td>101-1000</td>
<td>overlap similarity score</td>
</tr>
<tr>
<td>-o</td>
<td>21-50</td>
<td>overlap length</td>
</tr>
</tbody>
</table>
Table 4 - Pre- and post-filter contig and singlet counts for *K. brevis* EST assemblies using CAP3 overlap percent identity and overlap length cutoffs.

Pre-filter indicates raw CAP3 output sequences. Post-filter denotes sequences which, after trimming of their low phred quality ends, had lengths >= 50bp.

<table>
<thead>
<tr>
<th>Percent identity</th>
<th>Overlap length</th>
<th>Pre-filter</th>
<th>Post-filter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Contigs</td>
<td>Singlets</td>
</tr>
<tr>
<td>75</td>
<td>25</td>
<td>2638</td>
<td>5364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2651</td>
<td>5751</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2645</td>
<td>6690</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2668</td>
<td>8474</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2328</td>
<td>12033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>22351</td>
</tr>
<tr>
<td>75</td>
<td>50</td>
<td>2624</td>
<td>5688</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2638</td>
<td>6078</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2627</td>
<td>7007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2631</td>
<td>8753</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2297</td>
<td>12191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>22351</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>2635</td>
<td>5993</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2644</td>
<td>6368</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2638</td>
<td>7258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2627</td>
<td>8922</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2298</td>
<td>12198</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>22351</td>
</tr>
<tr>
<td>75</td>
<td>100</td>
<td>2689</td>
<td>6251</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2678</td>
<td>6606</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2662</td>
<td>7483</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2602</td>
<td>9103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2313</td>
<td>11938</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>22351</td>
</tr>
</tbody>
</table>
Table 4 - 6 The number and quality of BLAST hits and sequence complexity improves dramatically with pre-assembly filtering of *K. brevis* sequence reads.

The same initial population of reads was present for all assemblies but the top two rows (marked with an ‘*’*) were assembled after filtering polynucleotide sequences from reads, whereas no pre-assembly filtering was performed for the bottom three assemblies. For the two entries with a CAP3 identity threshold of 90 and a quality threshold of 20, the number of BLAST hits was 66% greater when pre-assembly filtering was used, and the average BLAST hit E-value rose from approximately $10^{-41}$ to $10^{-63}$. The average percentage of high-complexity regions in the annotated sequences also rose with pre-assembly filtering from 63% to 73%. Some fine-tuning was possible using the CAP3 identity threshold (the minimum identity between overlapping fragments in prospective clusters), and post-assembly using a simple quality threshold cutoff for inclusion into the database. However, there was an inverse relationship between the average E-value of the BLAST hits and the total number of BLAST hits – more stringent criteria produced less, better quality BLAST hits on average. The average complexity value was calculated as the average percentage of high-complexity regions output by the low-complexity filtering application *seg* across all permutations for the following parameters: window size (6,7,8), lower complexity threshold (0.8, 1.0, 1.2) and high complexity threshold (1.2, 1.4, 1.6). All data refer to contigs only (singlets were excluded from the analysis).

<table>
<thead>
<tr>
<th>CAP3 % identity threshold</th>
<th>Quality threshold</th>
<th>Avg quality</th>
<th>Avg length (bp)</th>
<th>Avg % high complexity</th>
<th>Average E-value</th>
<th>Number BLAST hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>90*</td>
<td>15</td>
<td>68.1</td>
<td>575.0</td>
<td>72.2%</td>
<td>-56.8</td>
<td>2477</td>
</tr>
<tr>
<td>90*</td>
<td>20</td>
<td>75.8</td>
<td>494.0</td>
<td>73.6%</td>
<td>-63.1</td>
<td>2356</td>
</tr>
<tr>
<td>75</td>
<td>20</td>
<td>68.7</td>
<td>575.2</td>
<td>62.5%</td>
<td>-57.9</td>
<td>1058</td>
</tr>
<tr>
<td>90</td>
<td>20</td>
<td>69.3</td>
<td>553.7</td>
<td>63.3%</td>
<td>-40.9</td>
<td>1415</td>
</tr>
<tr>
<td>95</td>
<td>20</td>
<td>70.6</td>
<td>545.1</td>
<td>64.4%</td>
<td>-25.4</td>
<td>1871</td>
</tr>
</tbody>
</table>

* Polynucleotide sequences trimmed from reads before assembly
Table 4 - Lack of consistency in Refseq annotation format hinders automated parsing of synonyms.
The variability of annotation formats precludes the use of grammar-based or rule-based parsers.

<table>
<thead>
<tr>
<th>Refseq ID</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC067218</td>
<td>N-acetyltransferase 2 (arylamine N-acetyltransferase)</td>
</tr>
<tr>
<td>BX641105</td>
<td>Alcohol dehydrogenase 1C (class I), gamma polypeptide</td>
</tr>
<tr>
<td>D90278</td>
<td>Carcinoembryonic antigen-related cell adhesion molecule 3</td>
</tr>
<tr>
<td>NM_001817</td>
<td>Carcinoembryonic antigen-related cell adhesion molecule 4</td>
</tr>
<tr>
<td>NM_000595</td>
<td>Lymphotoxin alpha (TNF superfamily, member 1)</td>
</tr>
<tr>
<td>NM_001816</td>
<td>Carcinoembryonic antigen-related cell adhesion molecule 8</td>
</tr>
<tr>
<td>NM_002764</td>
<td>Phosphoribosyl pyrophosphate synthetase 1</td>
</tr>
<tr>
<td>BC030975</td>
<td>Interleukin 1 receptor-like 1</td>
</tr>
<tr>
<td>BC071710</td>
<td>Interleukin 2 receptor, gamma (severe combined immunodeficiency)</td>
</tr>
<tr>
<td>D90070</td>
<td>Phorbol-12-myristate-13-acetate-induced protein 1</td>
</tr>
<tr>
<td>BC044792</td>
<td>Aminomethyltransferase</td>
</tr>
<tr>
<td>BC112192</td>
<td>HGF activator</td>
</tr>
<tr>
<td>BC045670</td>
<td>Cyclin-dependent kinase (CDC2-like) 10</td>
</tr>
<tr>
<td>D14838</td>
<td>Fibroblast growth factor 9 (glia-activating factor)</td>
</tr>
<tr>
<td>AY726568</td>
<td>Hypothetical protein LOC646870</td>
</tr>
<tr>
<td>L19315</td>
<td>Cholecystokinin A receptor</td>
</tr>
<tr>
<td>BM924710</td>
<td>Translocator protein (18kDa)</td>
</tr>
<tr>
<td>AF239668</td>
<td>Cholecystokinin B receptor</td>
</tr>
<tr>
<td>NM_000160</td>
<td>Glucagon receptor</td>
</tr>
<tr>
<td>NM_202002</td>
<td>Forkhead box M1</td>
</tr>
<tr>
<td>AB033337</td>
<td>M-phase phosphoprotein 1</td>
</tr>
<tr>
<td>NM_000379</td>
<td>Xanthine dehydrogenase</td>
</tr>
<tr>
<td>NM_004650</td>
<td>Patatin-like phospholipase domain containing 4</td>
</tr>
<tr>
<td>NM_004575</td>
<td>POU domain, class 4, transcription factor 2</td>
</tr>
<tr>
<td>NM_004288</td>
<td>Pleckstrin homology, Sec7 and coiled-coil domains, binding protein</td>
</tr>
<tr>
<td>BC012096</td>
<td>Arrestin 3, retinal (X-arrestin)</td>
</tr>
<tr>
<td>BF698898</td>
<td>Secreted phosphoprotein 1 (osteopontin, bone sialoprotein 1, early T-lymphocyte activation 1)</td>
</tr>
<tr>
<td>NM_002457</td>
<td>Mucin 2, oligomeric mucus/gel-forming</td>
</tr>
<tr>
<td>BC002419</td>
<td>Tar (HIV-1) RNA binding protein 2</td>
</tr>
<tr>
<td>AK160365</td>
<td>Rho guanine nucleotide exchange factor (GEF) 5</td>
</tr>
</tbody>
</table>
Table 4 - 8 Gene Ontology (GO) mapping files to external databases.
Multiple GO mapping files were used to construct a mapping table for finding annotation synonyms and corresponding IDs between different external databases.

<table>
<thead>
<tr>
<th>External database</th>
<th>No. entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG</td>
<td>69</td>
</tr>
<tr>
<td>ENZYME COMMISSION (EC)</td>
<td>3665</td>
</tr>
<tr>
<td>EGAD</td>
<td>35</td>
</tr>
<tr>
<td>GENPROTEC</td>
<td>236</td>
</tr>
<tr>
<td>go_component_ontology</td>
<td>1152</td>
</tr>
<tr>
<td>go_function_ontology</td>
<td>4922</td>
</tr>
<tr>
<td>go_process_ontology</td>
<td>18948</td>
</tr>
<tr>
<td>hamap</td>
<td>2795</td>
</tr>
<tr>
<td>INTERPRO</td>
<td>17312</td>
</tr>
<tr>
<td>METACYC</td>
<td>3937</td>
</tr>
<tr>
<td>MIPS</td>
<td>840</td>
</tr>
<tr>
<td>MULTIFUN</td>
<td>464</td>
</tr>
<tr>
<td>PFAM</td>
<td>8708</td>
</tr>
<tr>
<td>PRINTS</td>
<td>4460</td>
</tr>
<tr>
<td>PRODOM</td>
<td>1810</td>
</tr>
<tr>
<td>PROSITE</td>
<td>3720</td>
</tr>
<tr>
<td>REACTOME</td>
<td>2502</td>
</tr>
<tr>
<td>RFAM</td>
<td>491</td>
</tr>
<tr>
<td>SMART</td>
<td>943</td>
</tr>
<tr>
<td>SWISSPROT KEYWORD (SPKW)</td>
<td>713</td>
</tr>
<tr>
<td>TIGR</td>
<td>172</td>
</tr>
<tr>
<td>TIGRFAMES</td>
<td>4632</td>
</tr>
<tr>
<td>um_bdd_enzymeid</td>
<td>449</td>
</tr>
<tr>
<td>um_bdd_pathwayid</td>
<td>109</td>
</tr>
</tbody>
</table>
Table 4 - 9 Synonyms for ‘Putative 2Fe-2S ferredoxin’ derived from the external2go table ordered by relevance.

<table>
<thead>
<tr>
<th>Go description</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>electron carrier</td>
<td>thioredoxin-like 2Fe-2S ferredoxin</td>
</tr>
<tr>
<td>electron transport</td>
<td>Ferredoxin, 2Fe-2S type</td>
</tr>
<tr>
<td>electron transport</td>
<td>Ferredoxin [2Fe-2S], plant</td>
</tr>
<tr>
<td>electron carrier activity</td>
<td>Ferredoxin, 2Fe-2S type</td>
</tr>
<tr>
<td>iron ion binding</td>
<td>Ferredoxin, 2Fe-2S type</td>
</tr>
<tr>
<td>electron carrier activity</td>
<td>Ferredoxin [2Fe-2S], plant</td>
</tr>
<tr>
<td>iron ion binding</td>
<td>Ferredoxin [2Fe-2S], plant</td>
</tr>
<tr>
<td>electron transport</td>
<td>2Fe-2S ferredoxin, iron-sulfur binding site</td>
</tr>
<tr>
<td>electron carrier activity</td>
<td>2Fe-2S ferredoxin, iron-sulfur binding site</td>
</tr>
<tr>
<td>iron-sulfur cluster assembly</td>
<td>ferredoxin, 2Fe-2S type, ISC system</td>
</tr>
<tr>
<td>electron transport</td>
<td>2Fe-2S ferredoxin, plant-type, iron-sulfur binding site</td>
</tr>
<tr>
<td>electron carrier activity</td>
<td>2Fe-2S ferredoxin, plant-type, iron-sulfur binding site</td>
</tr>
<tr>
<td>iron ion binding</td>
<td>2Fe-2S ferredoxin, plant-type, iron-sulfur binding site</td>
</tr>
<tr>
<td>electron transport</td>
<td>[2Fe-2S]-binding</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>[2Fe-2S]-binding</td>
</tr>
<tr>
<td>electron transport</td>
<td>Rieske [2Fe-2S] region</td>
</tr>
<tr>
<td>2 iron, 2 sulfur cluster binding</td>
<td>2Fe-2S</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>Rieske [2Fe-2S] region</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>Methane/phenol monoxygenase/ferrodoxin-NAD+ reductase [2Fe-2S]-component</td>
</tr>
<tr>
<td>iron ion binding</td>
<td>Methane/phenol monoxygenase/ferrodoxin-NAD+ reductase [2Fe-2S]-component</td>
</tr>
<tr>
<td>metal ion binding</td>
<td>[2Fe-2S]-binding</td>
</tr>
<tr>
<td>2 iron, 2 sulfur cluster binding</td>
<td>2fe-2s cluster binding</td>
</tr>
<tr>
<td>electron carrier</td>
<td>2fe-2s electron transfer carrier</td>
</tr>
<tr>
<td>2 iron, 2 sulfur cluster binding</td>
<td>iron-sulphur cluster 2fe-2s binding</td>
</tr>
<tr>
<td>2 iron, 2 sulfur cluster binding</td>
<td>iron-sulfur cluster 2fe-2s binding</td>
</tr>
</tbody>
</table>
The largest deposits are the human (8.1 million sequences) and mouse (4.9 million sequences) EST collections, followed by thale cress *Arabidopsis thaliana*, cow, pig, and maize (all approximately 1.5 million sequences) and zebrafish *Danio rerio* (1.4 million sequences). There are 825 organisms with over 1,000 ESTs and 982 organisms with EST deposits of between 100 and 10,000 sequences.
The estuarine fish, *F. heteroclitus* is used in studies of the relationship between population genetics and environmental adaptation due to the extreme temperature range of its natural habitat: coastal shallows from New Foundland to the southern coast of the United States.

The term "red tide" or "Florida red tide" is used to describe a frequent algal bloom caused by the dinoflagellate *Karenia brevis* that occurs almost annually in Florida's coastal waters and in the eastern Gulf of Mexico.

*A. californica*, or the sea hare, is a large sea slug that inhabits the shallow coastal regions from Northern California to Baja California, feeding on seaweed and spawning in deeper water. *A. californica* is used extensively in studies of neurophysiology.

**Figure 4 - 2** The three EST species collections used in the development and testing of the myEST assembly and annotation platform: *Fundulus heteroclitus* (mummichog), *Karenia brevis* (red tide) and *Aplysia californica* (sea hare).
Figure 4 - 3 The synchronous interaction pattern of a traditional web application (top) compared with the asynchronous communication between server and client of an AJAX application (bottom) (Garrett 2005).
Figure 4 - The myEST plate display function shows marker and cloned sequence information for each 96-well plate in the collection.

Marker sequences are shown in green. The marker pattern used in this experiment was as follows: 1) Fixed markers in wells A05 and F09, 2) Plate specific marker in well corresponding to plate number (e.g., plate 1 would have a marker sequence inserted into wells H01 and A12, plate 2 would have markers in wells G01 and B12.
Figure 4.5 Cumulative assembly of *A. californica* sequence read libraries.

Cumulative addition of EST reads from *A. californica* results in a relatively steady increase in assembled sequences, suggesting that further addition of sequences will produce more unique sequences.
Figure 4 - Display of database statistics for the *F. heteroclitus* EST collection.

Unique sequences are defined as unique best BLAST hits – this will be less than the number of annotated sequences if some collection sequences have the same best hit target sequence. Unknown sequence have no BLAST hit. Quality values are phred quality scores.
Figure 4 - 7 Display of phred quality value distribution of contigs with no BLAST hits. Distributions can be selected by sequence type (singlet, contig, both) and by BLAST annotation type (annotated, unknown, both).
Figure 4-8 Two simple metrics for measuring the degree of coverage of contigs predicted from clusters of multiple aligned reads.

The read depth represents the maximum number of aligned read along the contig and the read coverage is a measure of the overall coverage by multiple reads supporting the sequence prediction along its entire length.

Read Depth = Greatest number of overlapping reads = 3

Read coverage = 1 read x (15% + 15%)
+ 2 reads x (20% + 30%)
+ 3 reads x 20%
= 0.30 + 1 + 0.60
= 1.90
Figure 4 - 9 Model-View-Controller implementation of the Funnybase 2.0 web interface.

The logical division between the three functions contributes to ease of maintenance and customization. The Model portion encapsulates the data model implemented in the databases and the CGI scripts used to retrieve the data. The View is the combination of an HTML template and a cascading style sheet (CSS) file which create the layout and style of the pages. The Controller is the Javascript file which is also loaded onto the client and provides the logic behind the page functions by accepting user inputs, translating them into AJAX requests from the server and loading any retrieved data into the page.
The workflow page can be viewed by selecting File → Open → Workflow from the menu. This page is used to configure and run workflows and monitor job progress. The different tabs of the workflow page provide the following functions: A) **Run**: Any workflow may be started and stopped and its progress is graphically displayed. Information on the status of any stage in the workflow can be viewed by placing the mouse over the workflow icon. B) **Workflows**: Each database is composed of one or more workflows. Clicking on a workflow selects its stages, which can be viewed in the Stages tab. C) **Stages**: The names and descriptions of each stage in the selected workflow are listed in order of execution. Clicking on a stage selects its application, which can be viewed in the Applications tab. D) **Applications**: The parameters used in any particular stage of the workflow can be altered by editing the arguments for the application; clicking on the listed arguments allows in-line editing, which is saved to the database. The next time the workflow is run, the new arguments will be used for this application.
Conclusion
The objective of this research was to improve EST database annotation using external databases. We have shown in Chapter 2 that a significant portion of genes in UniGene, the largest publicly available EST database, have no detectable counterparts in the other UniGene databases or have non-functional annotation. Compounding the problem of incomplete sources for sequence comparison is what may be an unexpectedly high misannotation rate. By an exhaustive BLAST of Swiss-Prot against itself combined with fuzzy matching, we showed that BLAST hits between functionally unrelated proteins can occur at surprisingly high rates at very low E-values. This suggests that annotation based on ‘best BLAST hit’, particularly with the commonly-used E-value cutoff of $10^{-5}$, may be prone to greater error than would be expected based on the literal definition of E-value: the expected number of chance hits between two unrelated random sequences with a particular E-value score in a particular database. To deal with this problem, we proposed an incremental mismatch probability based on the actual BLAST hit distribution of correct and incorrect matches as an indicator of the likelihood an annotation is correct. Using an incremental mismatch probability threshold of 0.05 reduced the number of incorrect annotations by half in our sample of ‘lowest E-value’ BLAST annotations of 1,000 human UniGenes against three other mammalian UniGene databases (mouse, dog and rat) and Swiss-Prot. There was also some loss of correct annotations due to false negatives but the Type II error rate was reasonably low and could be optimized by
using a higher mismatch probability. However, this emphasis on reducing the Type I error rate (of accepting incorrect annotations) is in accordance with the need to avoid cascading errors due to interdependence in database annotations.

In an effort to classify orthologous groups of proteins by their motif signature – the combination and order of protein motifs they share – we have shown that naïve Bayes and decision tree models can predict sequence function with a high degree of accuracy. The InterPro Motif Annotation Tree (IMAT) algorithm categorizes InterPro motifs and uses decision trees to model protein function as a directional signature of motifs. The advantage of IMAT is the relative ease of interpretation of the biological dynamics underlying Decision trees. Because this ‘glass box’ model of protein function is more intuitive and easily understood than other machine learning models, it has a dual use: the classification of genes and gene fragments, such as ESTs, and also as a conceptual model of the relationship between protein function and structure. The IMAT library will be provided as an open-source annotation resource for the genomics community.

In Chapter 4, we discussed how EST databases continue to play an important role in the genomic characterization of non-model organisms whose genomes have not been sequenced, and are also vital for the correct interpretation of genomic sequences. With the accelerating increase in the number of sequenced genomes, non-model
organisms are increasingly important as convenient platforms for analyzing particular biological systems and as interspecies reference points. However, errors rates at the various stages in EST assembly and annotation can undermine the value of EST databases. This release (version 2.0) of the *Fundulus heteroclitus* database (Funnybase) was built using the open source myEST software platform, a suite of Perl applications designed for improving EST assembly quality control and annotation. The myEST platform also includes a web interface with enhanced data accessibility and visualization, and customizable high-throughput genomics analysis workflows. It is intended as a free and customizable resource for EST and other sequence projects.