

# **Development and comparison methods for ab-initio classification of 5' ESTs**

T.B. Bair, T.E Scheetz and T.L. Casavant

## **Introduction/background**

Since the sequencing projects of the human, mouse and rat genome are all nearing, or in a state of completion, considerable effort has shifted to the annotation, and use of these genomic resources. These resources have revealed in many cases the weakness of common abstractions dealing with gene structure and function. The notion of a gene has undergone considerable scrutiny and through the use of gene prediction programs and EST sequencing projects, several preliminary estimations of the number of genes have been proposed [Aparicio]. These methods of finding genes have been developed to a high degree. Gene prediction methods have been implemented with specificity and sensitivity of many techniques measured in the high 90% range for test data sets [Rogic]. The limitation though is the number of mis-predicted genes, with even a small error rate the number of mis-predictions is large over the complete genome. The combined with the tendency for gene prediction programs to predict genes with structure similar to their training sets. These questions indicate that considerable doubt remains as the exact transcriptome even with the complete genome.

Prediction methods will never equal the specificity of experimental methods, particularly when looking at alternative splice forms, extremely short or long genes, or other genes with structures which are not currently in the typical gene prediction test sets. It is unlikely that any gene prediction program would soon be able to accurately identify all of the alternative splice forms that would be biologically significant, particularly with splice forms that are from a limited range of tissues. Yet these genes and splice forms are exactly the ones that are most interesting; the genes and splice forms which are not yet defined or adequately studied.

A method that has been used to address the limitations of gene prediction programs is large-scale EST sequencing. These are efforts that attempt to find a short, usually 3' sequence, for most genes in a particular tissue or cell type. These projects begin to address the question of how many genes are actually present. With enough redundancy in the ESTs sequenced and tissues examined, questions of the total number of transcripts also can be addressed. However, EST libraries tend to focus primarily on the 3' end with a secondary emphasis on the 5' end. Very few ESTs are intentionally from the middle of RNA transcripts. This limits the assessment of alternative splice forms in the coding region and is misled by alternative polyadenylation. Also, EST libraries are not without artifacts which may make them just as error prone as gene prediction methods.

Projects such as RIKEN [FANTOM] and MGC [Strausberg] try to find and fully sequence an mRNA clone from every gene. These projects can begin to completely characterize the transcriptional diversity that is present within a variety of tissue types. By sequencing at a very high quality standard, and characterize the complete gene, these projects can eliminate much of the error that is present in EST projects, and leave little question that the gene in the database is probably a real transcript.

The problem in full-length sequencing projects is identifying clones with full-length genes from which the complete sequence can be found. It does little good to

sequence a clone that contains only a portion of the full gene sequence. The difficulty lies at two points: assessing the clone with a minimum of cost and finding genes which have not yet been characterized. It is relatively straightforward to assess a gene if you have the complete sequence or if you are aligning a portion of that sequence to a known annotated gene. It is much more difficult to make that assessment with only a portion of that gene (EST) or if the goal is to find genes for which no known homologue exists.

## Methods

An important factor in the development of ab-initio prediction methods is the selection of a test set. Inaccuracies or biases in the test set influence the quality of the prediction as well as bias assessment of the accuracy of the program. Often the challenge in development of a test set is balancing the quality of the test set with the quantity available at a given quality standard. Over the course of the development of this program we have used two methods to generate test data sets. This evolution was due to the later availability of a more realistic test set that was obtained by the categorization of EST sequences. This is the preferred approach since the test data accurately reflects the quality, length and errors present in EST data. This does have the drawback of introducing some bias. The bias is that the selected genes will resemble the ESTs that have already been selected; this may be an acceptable compromise.

The other test set is a “synthetic” EST that was generated from truncated full-length UniGene sequences. The advantage of this test set is that it generates more sequences to test with and this data was available to us first.

### Generation of synthetic EST test data

To generate a test set of EST like sequences we used annotated full-length genes that were present in the mouse UniGene set [Schuler]. The first step was to collect annotated full-length genes from the unigene set. The set was filtered to contain only the 21,000 clusters which have a full length annotated gene. From the full-length genes we then took a random portion of the sequence approximately 500bp long. The target length is a parameter but the actual length is controlled by the location of the start and the availability of sequence after the selected start with the actual sequence being  $\leq 500$ bp. With a subsequence of the original gene obtained randomly, the subsequence is then categorized. The location of the returned portion, either CDS, having short /5' UTR ( $< 50$ bp), or having long UTR ( $> 50$ bp) is annotated.

The goal of this exercise was to ensure that the data was similar to what would be found in a perfectly sequenced EST library. It is important to note that this test set probably does not contain sequencing errors which, would be expected in normal EST sequences, thus prediction programs would be expected to perform better on this set than with ‘real’ EST data. The synthetic ESTs (synEST) were categorized based on the position of the cuts relative the annotated start position. The categories were: good-long UTR, good-short UTR, and bad. This categorization is based on the length of the UTR and if the synEST overlapped with the annotated start.

### Collection of annotated ESTs

For the creation of this dataset ESTs, from an actual high throughput sequencing project, are selected that are homologous to known genes. These ESTs either overlap to

the annotated start codon or not. If the EST overlaps the annotated start they are selected as full-length, if not they are annotated as not-full-length. These ESTs are obtained from our collaborator Dr. Bento Soares lab. We are fortunate to collaborate with Dr. Bento Soares lab which is able, through a variety of methods [Bonaldo, Soares], to produce high quality clone libraries which are enriched for full-length clones, size fractionated, and enriched for novel sequence. These high quality libraries provided us with a reasonable number of ESTs to test or prediction methods. In addition, the Soares lab full insert sequenced a number of clones that we identified as potentially full length. The importance of this set is to mirror the variability of real ESTs and to determine how well the programs are at identifying clones whose 5' ESTs are from the beginning of novel genes.

It is important to note that the homology based assignment of ESTs may in fact be incorrect in some instances. This is due common to all homology methods such as gene families, incorrect annotation and others. This may in turn incorrectly assign other ESTs which are being evaluated.

The ESTs were assigned by blasting the ESTs to databases which contained annotated full length genes. The databases that were used were Swiss-Prot [Boeckmann], mouse MGC [Strausberg], human MGC [Strausberg], RIKEN [FANTOM], RefSeq (mouse)[NCBI], and HUGE [Kikuno]. If the EST had a match to one of these databases with a 90% match over 300bp long the EST is selected for annotation either as full-length or not full-length.

#### isFulllength algorithm

Figure 1 outlines the algorithm used for ab-initio prediction. The method attempts to classify an EST based on a number of parameters. These parameters include the location of putative start codons, the location and relationship of stop codons to the putative start, fidelity of a Kozak consensus region near the start, length of open reading frame (ORF) after the putative start, and the effects of a frame shift on most of these parameters. The effects of a frameshift are a very important consideration since EST has highly variable error rates [Hillier].

The isFulllength algorithm will assign a given EST in to a clone class. This clone class is a description of all the features that might be relevant to assessing the ESTs full-lengthness. This classification is an over-specification, that is many classes will be considered full-length. The clone class is somewhat hierarchical but the full-lengthness of the tree does not segregate to one location of the tree rather it is a series of terminal nodes with a few branches being assessed as all potentially full-length.

```

Foreach ATG in the 5'EST
  A. Upstream and downstream of the ATG, identify:
      1.Number of stops in -frame
      2.Number of stops in all frames
      3.Number of frames containing stops
      4. Length of hypothetical ORF and UTR
  B. ATG_score =  $\Sigma$  weighted differences from an ideal vector of features
  C. if (min(ATG_score) < ATG_threshold) then
      Initially classify as full-length
  D. else-if (min(ATG_score) > CDS_threshold)
      Initially reject as full-length
  E. else
      Initial classify as UTR
  F. Annotate surrounding features

```

**Figure 1. Method *isFulllength* used to determine weighted sum of ATG-derived features**

### Manual classification of results

The *isFulllength* algorithm at its heart only describes the EST. It is a secondary process to classify the output from that algorithm. Initially, this was done by using an EST test data set similar to the synEST dataset. After categorizing the synESTs, percentage yield vs percentage error calculations were done for each clone class. The ones that provided a maximum return with minimal error were chosen as representing full-length ESTs. The clone classes that were chosen as the best predictors are listed in table 1.

**Table 1: Listing of clone classes that are indicative of a full-length clone based on EST sequence**

#### High Confidence

KG1E.ATG.L.zSL.ORFr0FS.  
 K0E.ATG.L.npSL.ORFr0FS.  
 K0E.ATG.M.npSL.ORFr0FS.  
 K1E.ATG.L.pSL.ORFr0FS.  
 K1E.ATG.M.pSL.ORFr0FS.

#### Lower Confidence

K0E.ATG.L.npSL.ORFr1FS.  
 K0E.ATG.L.pSL.ORFr0FS.  
 K0E.ATG.L.pSL.ORFr1FS.  
 K0E.ATG.M.npSL.ORFr1FS.  
 K0E.ATG.M.pSL.ORFr0FS.  
 K0E.ATG.M.pSL.ORFr1FS.  
 K1E.ATG.L.npSL.ORFr0FS.  
 K1E.ATG.L.npSL.ORFr1FS.  
 K1E.ATG.L.pSL.ORFr1FS.  
 K1E.ATG.M.npSL.ORFr0FS.  
 K1E.ATG.M.npSL.ORFr1FS.  
 K1E.ATG.L.pSL.ORFr1FS.  
 K1E.ATG.M.npSL.ORFr0FS.  
 K1E.ATG.M.npSL.ORFr1FS.  
 K1E.ATG.M.npSL.ORFr0FS.

K1E.ATG.M.npSL.ORFr1FS.  
K1E.ATG.M.pSL.ORFr1FS.  
KG1E.ATG.L.npSL.ORFr0FS.  
KG1E.ATG.L.npSL.ORFr1FS.  
KG1E.ATG.L.pSL.ORFr0FS.  
KG1E.ATG.L.pSL.ORFr1FS.  
KG1E.ATG.M.npSL.ORFr0FS.  
KG1E.ATG.M.npSL.ORFr1FS.  
KG1E.ATG.M.pSL.ORFr0FS.  
KG1E.ATG.M.pSL.ORFr1FS.

#### Decision tree classification of results

The second method that was used to classify results was a decision tree based approach. The first step in this process was to create a parser that extracted the ATG score, ATG position and the kozak consensus sequence position and finally the clone class description. This clone class description was a concise description of what categories *isFulllength* put the clone into. A typical clone class description might look like "K0E.ATG.L.npSL.ORFr0FS" which would indicate that the kozak would have 0 errors "K0E", that there was an ATG "ATG", that the ATG was in the left third of the EST "L" and number of frames with stops upstream of the start "npSL" has an open reading frame "ORF" with 0 frame shifts. Table 2 gives a complete description of all clone class descriptions.

**Table 2: Detailed analysis of encodings present in the isFullLengh clone class description**

Encoding	Description
K0E	Kozak Good 0 errors from consensus
K1E	Kozak found with 1 change from consensus
KG1E	Kozak found 1 error or less from consensus but not with the best ATG
zSL	Zero Stops upstream of ATG
pSL	Number of stops left of ATG > the average in all three frames
npSL	Number of stops upstream of ATG is higher than expected
R	ATG is in 3' 1/3 of EST
M	ATG is in the middle 1/3 of EST
L	ATG is in the 5' 1/3 of EST
noORF	No open reading frame over 100 bp long
ORFr0FS	ORF present with no induced frame shifts
ORFr1FS	ORF present but is longer with an implied frame shift
nSDistrMed	Number of total stops medium
nSDistrHi	Number of total stops high
nSDistrLo	Number of total stops low
CDS	Est likely to be all coding
nS0	
Reject Short	EST is too short to be considered
ATG	ATG likely to be found
UTR	Likely 5' UTR

The parser takes the page long detailed description returned by *isFulllength* and returns the ATG score, ATG position and a delimited list of terms found in the clone class description. This is in turn sent to a program which uses the perl module AI-DecisionTree-0.06 which is freely available at [www.cpan.org](http://www.cpan.org). This module is trained by showing this module the *isFulllength* outputs from known full-length and known non-full-length genes in a training mode. Then the decision tree is saved and tested on EST data that the module has not yet been exposed to. From the results of this test measures

are made as to the effectiveness of the decision tree at predicting the type of EST given the output of *isFulllength*.

### Assessment of results

The prediction methods were assessed by several simple statistical measures. Sensitivity = True\_Positive/ (True\_Positive+False\_Negative). Specificity = True\_Negative/(True\_Negative+False\_Positive). Also used is predictive positive value (PPV) which, is a measure of the likelihood of a selected clone being full-length (PPV = True\_Positive/(True\_Positive+False\_Positive)). Finally, the term “correct” is used to denote selections that are correct (% True\_Negative + % True\_Positive). We have found all these measures essential in determining the effectiveness of the ab-initio prediction.

## Results

### Assessment of the agreement in the EST annotation

Our collection includes over 90,000 5’ ESTs which upon clustering [Trivedi] form approximately 28,000 unique sequences. These sequences are categorized by blasting them vs. a number of specialized databases and looking for an overlap between the EST and the annotated start position of the full-length transcript. The result of this categorization is shown below along with indications on agreement between the databases; particularly the agreement with the mouse MGC set which is thought to be a particularly high quality database.

**Table 3: This table shows the number of ESTs from our set which is annotated either good (full-length) or bad based on the probable overlap of the ATG with the annotated start site. The overlap and agreement with Mouse MGC is based only on the full-length sequences for which we have local EST copies in our system and is intended as a measure of overlap between the databases and the amount of discrepancies between the annotations.**

Database	Number full-length	Number not full-length	Overlap with M-MGC	Agreement with M-MGC
Swiss-prot	1616	3267	35.6	99.1
Mouse MGC	2155	3525	100	100
Human MGC	504	652	44.8	98.2
Riken	1475	1799	44.3	97.4
RefSeq	2037	4209	37.9	99.1
HUGE	61	395	17	97.8

The following table shows the agreement of the two ab-initio methods with the different database predictions. While there are slight variations in the specificity and sensitivity from each database the overall trend is very clear. It is likely that the variation is due to either mis-annotation in the databases or a incorrect homologous association.

**Table 4: Sensitivity and Specificity of the two methods, Manual: where the clone classes were selected manually and Decision Tree: where the clone classes were selected by a decision tree**

method, where the methods were compared two various databases. The mouse MGC results for the decision tree method should be regarded with caution since this set was used in the training of the decision tree. The fact that the results were not 100% is indicative of the complexity of the data.

Database	Sensitivity Manual method	Specificity Manual method	Sensitivity Decision Tree	Specificity Decision Tree
Swiss-prot	5.8	99.1	50.5	77.9
Mouse MGC	5.9	99.2	46.9	83.1
Human MGC	4.3	99.5	46.1	79.5
Riken	3.9	99.3	38.6	84.9
RefSeq	5.9	99.2	49.7	78.8
HUGE	9.8	99.0	54.1	78.9

The next table shows the total number of predictions made that would be examined further, with no evidence contrary to this assessment, (other predictive methods) these ESTs would be selected for further sequencing. These numbers are important due to the goal of our prediction methods which are to keep a full-length sequencing pipeline supplied with the best possible choices of EST. Some methods and databases, while reliable, do not have enough sensitivity to fully supply the pipeline with new candidates on a continuing basis.

**Table 5: The total number of clones that would be selected by a particular database or method is listed here. The fact that the decision tree method can find >6000 clones which it considers full-length with an accuracy of ~78% indicates that it will correctly identify 4-5000 clones, far more than could be identified by homology to RefSeq or any other database.**

Database/Method	Number predicted as good (from 27913)
Manual method	499
Decision Tree	6094
Swiss-prot	1614
Mouse MGC	2152
Human MGC	503
Riken	1473
RefSeq	2034
HUGE	61

Examples of ESTs that were found that could not have been found by homology based methods.

In a retrospective study of 396 confirmed full-length clones. These were clones that were selected by a variety of methods, the clones were fully sequenced. These clones were selected both by homology and ab-initio prediction methods. One particular bias that must be noted was the intentional avoidance of clones already in the mouse MGC set. This was due to the eventual goal of depositing the fully sequenced clones into the MGC collection. Some clones were selected that were in the mouse MGC set. This is either due to a lag in database updates or through later inclusion of the clone in the MGC set which was examined here.

While only 10% of the clones presented here were selected by ab-initio methods alone, in the future this number will inevitably rise. This is due to both the early reliance of these methods (they are undoubtedly more specific and should be used first) and to the eventual depletion of the clones that can be picked by homology methods.

**Table 6: Retrospective study of clones selected and fully sequenced. The clones were selected by a variety of methods beyond what are described in this paper. This table attempts to show how often a particular database or ab-initio method would select these known correct clones. It should be noted that the lower numbers from the mouse MGC set are artifactual due to an intentional avoidance of overlap with this set.**

Prediction type	Predicted as full-length	Not predicted as full-length (Bad or no prediction)	Percent correct
Ab-initio manual	36	361	9.1
Ab-initio decision tree	202	195	50.9
Ab-initio methods only	40		
Swiss-prot	109	287	27.5
Mouse MGC	66	306	17.7
Human MGC	59	317	15.7
Riken	178	198	47.3
RefSeq	141	229	38.1
HUGE	17	353	4.6

## Conclusions

The use of this algorithm has allowed us to begin to find full-length genes which have no homology to any previously known gene. This is important, if you only base your search for full-length genes you will: first probably only find a subset of the known genes and second, be unable to add new genes to that set.

Our goal in this process, as with most prediction processes, is to maximize sensitivity and specificity. However, due to the nature of the processes we are working on, we would be willing to compromise on specificity. Because full-insert sequencing is decreasing in cost, and the value of novel full-length sequences our preferred error mode is to relax specificity. This is why the decision tree additions to the process were so critical. They allowed us to find many more ESTs, most of which are full-length to sequence.

Both the manual selection of clone classes and the use of decision trees on the dissected clone classes are valid approaches. If the desired goal is to have near absolute certainty that the selected clone is truly full-length this is the better method. However, the sensitivity of this method is lacking. The decision tree method is much more balanced, yet its specificity is still acceptable.

This is still an open area of development. We will continue to develop the methods presented here by experimenting with other machine learning approaches, and by using complimentary prediction methods which will include aligning the ESTs to

genomic sequence. It is an advantage that the methods presented here will work for most mammalian EST projects; however the additional information that can be provided by the mouse, rat and human genomes in conjunction with the EST sequence cannot be ignored.

## References

Aparicio S How to count human genes *Nature Genetics* 25:129-30 2000.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., Oâ€™Donovan C., Phan I., Pilbout S. and Schneider M. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31:365-370(2003).

Bonaldo MF, Lennon G and Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 1996 Sep;6(9):791-806.

FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs *Nature* 420:563-573, 2002

Kikuno R., Nagase T., Waki M. et al., HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.* 30, 166-168, (2002).

NCBI. The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 17, The Reference Sequence (RefSeq) Project. Available from <http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books>

Schuler G.D. et al. A gene map of the human genome. *Science* 274: 540-546. 1996.

Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci U S A* 1994 Sep 27;91(20):9228-32

Strausberg RL, Feingold EA, Klausner RD, Collins FS. The Mammalian Gene Collection. *Science*, 1999, 286, 455-457.

Trivedi, N., J. Bischof, S. Davis, K. Pedretti, TE Scheetz, TA Braun, CA Roberts, NL Robinson, VC Sheffield, MB Soares, TL Casavant. Parallel creation of non-redundant gene indices from partial mRNA transcripts. 2002 *Future Generation Computer Systems* 18(6):863-870.