UPTEC X 04 047
DEC 2004

ISSN 1401-2138

DANIEL ANDERSSON

Finding novel full length genes using EST data

Master's degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 04 047	**Date of issue 2004-12**

Author

Daniel Andersson

Title (English)

Finding novel full length genes using EST data

Title (Swedish)

Abstract

A method using data from expressed sequence tags in gene predictions was evaluated by comparing Unveil, which had previously been modifed, with Genscan and GeneID. The data set was the complete human genome. Tests were also performed to optimize the method and to check whether the programs discovered different genes. The method increased the accuracy of Unveil greatly. However, the modified Unveil is not as good as the other programs.

Keywords

EST, gene prediction, Unveil, evaluation, human genome

Supervisors

Helgi Schiöth
Department of Neuroscience, Uppsala University

Scientific reviewer

Robert Fredriksson
Department of Neuroscience, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **43**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Finding novel full length genes using EST data

Daniel Andersson

Sammanfattning

Det humana genomet blev sekvenserat år 2001. Nästa steg är att identifiera och lokalisera generna vilket är ett mödosamt och tidsödande arbete. För att förkorta tiden så har olika dataprogram som förutsäger gener utvecklats. Ett av de vanligaste programmen är BLAST. Det använder redan kända gener för att identifiera tänkbara nya gener. Denna metod kommer dock hitta endast en bråkdel av alla gener. En annan metod som förlitar sig på hur gener tenderar att se ut används därför flitigt. Det bästa vore en kombination av de två. Dr Helgi Schiöths grupp på Institutionen för neurovetenskap har uppdaterat ett program som heter Unveil så att det använder delar från bägge metoderna.

Innan en specifik metod används så bör den utvärderas. Detta har normalt gjorts på data som inte stämmer med verkligheten. De tidigare testerna har därför inte gett ett bra mått på hur bra programmen är. För att få ett bra mått så bör hela det mänskliga genomet användas. Anledningen till att detta inte har gjorts tidigare är att den kompletta uppsättningen mänskliga gener inte är känd och därmed finns det inget att jämföra förutsägelserna med. Detta är fortfarande sant, men databaserna över våra gener börjar dock bli tillräckligt bra för att kunna användas som facit.

Det nya Unveil utvärderades genom att jämföra det med det gamla. Resultatet visar att ändringarna som gjordes i Unveil ledde till en klar förbättring. Det nya Unveil är dock inte lika bra som GeneID och Genscan.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet december 2004

Contents

List of Tables

List of Figures

1 Introduction

The first draft version of the human genome was released in 2001 by the International Human Genome Sequencing Consortium [1]. However, only knowing the sequence of the human genome is almost the same as knowing all the letters in a book but not the words or sentences, i.e. almost worthless. What we are really looking for are the human genes. Knowing the genome sequence might not tell us how many or where the genes are located but it is a big help as it allows the use of computer programs in the search for the genes.

Even today, a couple of years after the first programs were made, none is perfect. This means that it is interesting to compare all the different programs in order to find their different strengths and weaknesses. Knowing those make it easier to pick the best program for the task at hand.

The first comparative study of programs that predict gene structures were published in 1996 by Moises Burset and Roderic Guigo [2]. This study constructed a set of genomic sequences from vertebrates that contained one gene per sequence, and hardly any non coding DNA. The most used programs of the time was then tested to see how well they performed on this set. The study, even though it is quite valuable as it establishes a procedure for comparing programs, suffers from a problem. It is performed on a unrealistic set of sequences. Normally you do not have just one gene per sequence and almost no non coding DNA.

Another study that used sequences from mammalian species, instead of from the complete vertebrate subphylum, was published in 2001 [3]. It went a little deeper than the study by Burset and Guigo as it looked at different factors and how they influenced the results. But, the set used still suffered from the above mentioned problems; only one gene per sequence and hardly any non coding DNA.

One reason most studies decide to do the analysis on a set that are far from realistic is that it is hard to find a good annotation that covers a large stretch of genomic DNA. Guigo *et al*, solved one of the problems in a later study. They added random DNA to both ends of each sequence in the test set, thus adding lots of non coding DNA. As expected the performance of all programs dropped significantly [4].

Enough human genes have been found, or at least predicted with a high degree of certainty, that a comparative study of multiple programs over the complete human genome is now viable. This study will give a more accurate assessment of how well the programs perform in a real situation. The four programs included in the study are Unveil, Unveil + EST, Genscan, and GeneID. Unveil + EST is a modified version of Unveil that uses information gathered from ESTs to increase the accuracy of its predictions.

2 Theory

2.1 Expressed Sequence Tags

If all the mRNA in the cell could be sequenced then it would be a valuable source for gene prediction programs and it would greatly improve their accuracy. However, mRNA is very unstable and breaks down easily outside the cell. This makes sequencing hard. The problem can be solved by converting the spliced mRNA to complementary DNA (cDNA) with the help of enzymes. The expressed sequence tags (ESTs) are then made by sequencing a few 100 nucleotides from both ends of the cDNA. Later the ESTs can be used to identify genes in a similarity search (see 2.2.1).

2.2 Finding Genes

Finding genes is a complex task as in order to actually find a gene you need to predict the complete gene structure, i.e. all the exons, start and stop codons, correctly. This might not sound like a tough thing to do but even a single incorrectly predicted nucleotide means that the entire gene is incorrect.

There are two primary ways to finding genes [5], similarity and probability search. Both ways have advantages and disadvantages and quite a few programs try to incorporate things from both.

2.2.1 Similarity Search

A similarity search uses known gene sequences, ESTs, and cDNA to look for new genes [5]. It exploits the fact that genes can look alike. When doing a similarity search you compare your sequence, using for example the BLAST algorithm, against a database of known genes from the same specie or from related species. If a good enough match is found it is a high probability that a gene exists. However, nothing will be found if the database does not contain a similar sequence which means that this approach is limited when it comes to finding new genes. Another problem is that even if you find a position in the sequence with a good match it is very hard to locate the exact exon-intron boundaries as the match probably includes gaps.

2.2.2 Probability Search

Programs that use principles based on probability search often use a Markov model to find genes. A Markov model consists of different states and each state has a certain possibility to emit, or generate, a character or a string of characters. The models are classified on the basis on how many previous states a specific state is dependent on when it comes to the probability to emit characters. These classes are called k-order Markov model, see figure 1, where k is the number of previous states the k+1 state depends on.

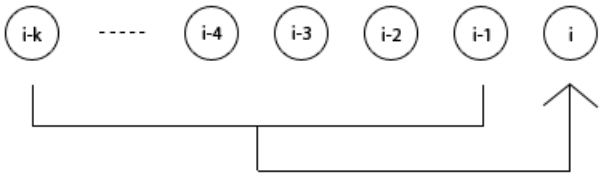


Figure 1: k-order Markov model

The most common Markov model, not necessarily in gene prediction, is the Hidden Markov Model (HMM) which states are not dependent on the previous states. In certain applications, such as sequence alignment, this property is useful but in gene prediction it causes problems as there is a well documented codon bias and thus the positions in the codon is dependent on each other. It is therefore necessary to model the dependency between the first, second, and third position in the codon in a different way if a Hidden Markov model is used. Unveil, see 2.3.1, has solved this problem by building a HMM which is constructed in three layers with 4-16-64 states. Parts of this HMM can be seen in figure 2.

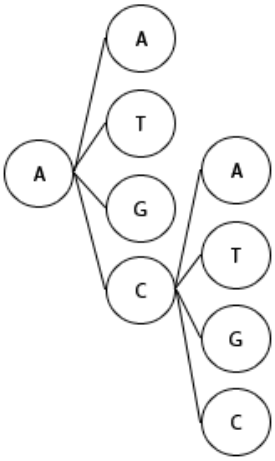


Figure 2: Part of Unveil's Exon HMM

Instead of the complex structure used by Unveil the dependency between the different positions in the codon can be modelled by a 2-order Markov model. However, it has been shown [6] that a 5-order Markov model models the different dependencies that exist the best. Genscan, see 2.3.4, therefore uses such a model for the non-coding regions. Using a higher order Markov model must be weighed against the need for more sequences to train the model [5]. Another common Markov model is the generalized HMM where each state uses different submodels, e.g. weight matrices and neural networks [7]. Each state can also have its own length distribution and emit more than one character.

The probability of finding genes are dependent on the amount of GC. Therefore the newer programs tend to use different models for different GC content. Most uses the GC ranges introduced by Genscan [8].

2.3 Programs Tested

All the tested programs were run on SweGrid, a grid of computers operated by the technical colleges in Sweden. The computers in the grid are Pentium IV 2.8 GHz with 2 Gb of RAM.

2.3.1 Unveil

Available at *http://www.tigr.org/software/Unveil/*. Unveil is a 286-state Hidden Markov Model [9, 10], based on the Veil design [11]. The HMM is split up into eight submodels, each a small HMM in itself. Each submodel is used to predict a certain feature of the gene and is later merged into a predicted gene. The different features are start codon, stop codon, exon, donor site, acceptor site, intron, frame shift and intergenic region. The output produced follows the GFF standard [12]. Unveil requires quite a lot of memory to run, around 2 Gb for a sequence of 700 kilobase-pairs (kbp). This fact makes Unveil unusable for predicting genes in large genetic material. However, Mattias Bäck has, as a part of his master thesis at the Department of Neuroscience Uppsala University, updated Unveil so that it reads a part of the sequence and then the next part and so on [13]. This approach means that Unveil does not have an upper limit when it comes to sequence length. Regardless of this upgrade the memory limit still applies and Unveil can not handle parts larger than 700 kbp. Even if there is no upper limit for the sequence length, running the program on really long sequences are not recommended as the program is rather slow.

2.3.2 Unveil + EST

Not available [13]. Unveil + EST is a modified copy of Unveil that uses information gathered from expressed sequence tags (ESTs) to improve its predictions. It works by increasing and decreasing the probability that an exon exists depending on how many different ESTs overlap the potential gene. The memory requirements are the same as for the unmodified Unveil but the running time has increased a little due to the extra checks for overlapping ESTs.

2.3.3 GeneID

Available at *http://www1.imim.es/software/geneid/*. GeneID works by first predicting splice sites, start and stop codons and then creating exons from these predictions [14]. The last step is constructing the optimal gene from the constructed exons. The output produced is either in GeneID's own output or GFF [12]. GeneID can be used on any sequence length, as it seems to read the sequence in parts in a similar way to the modified Unveil. When it comes to speed it is hard to be faster than GeneID. It managed to read and predict genes on the human chromosome 22 in around 7 minutes.

2.3.4 Genscan

Available at *http://genes.mit.edu/license.html*. Genscan is based on a generalized HMM [8]. In an generalized HMM each state can emit a sequence of any length and can use different models. Genscan predicts more than just the gene structure. It also finds poly-A tails, cap sites, TATA-boxes and a few other features. The program includes suboptimal exons in its predictions, and lets the user decide what the cut-off should be. The default values were used for the predictions in

this report. As Genscan tries to read the entire sequence into memory it has problems with longer sequences. Speed-wise it is a little slower than GeneID.

2.4 Comparing the Programs

Statistics were collected on three different levels: nucleotide, exon, and gene. No other data was considered when comparing the different programs.

2.4.1 Nucleotide Level

On the nucleotide level the statistical measurements introduced by Moises Burset and Roderic Guigo [2] were used. Each predicted nucleotide will either be coding or non–coding, and the same holds true for all nucleotides in the annotation. By comparing the status of a specific nucleotide in the prediction with its status in the annotation it can be classified in either of four groups

TP Nucleotides that have correctly been predicted as coding.

TN Nucleotides that have correctly been predicted as non–coding.

FP Non–coding nucleotides that have incorrectly been predicted as coding.

FN Coding nucleotides that have incorrectly been predicted as non–coding.

The two most commonly used statistical measurements are sensitivity (Sn) and specificity (Sp).

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

The normal definition (2) of specificity is not good as the number of non–coding nucleotides in human DNA is much larger than coding ($TN \gg FP$), which means that $Sp \simeq 1$ almost all the time and thus does not provide any useful information. Instead, the definition used in gene prediction is (3).

$$Sp = \frac{TP}{TP + FP} \tag{3}$$

Neither of these measurements summaries the global accuracy very well. It is quite easy to reach a high sensitivity, just predict all nucleotides as coding. The same is true for specificity, predict very few coding nucleotides. If a single value is to be used to compare how well programs do at the nucleotide level it needs to take both sensitivity and specificity into consideration at the same time. The most straightforward measurement is just to take the mean (4) of the sensitivity and specificity.

$$SnSp = \frac{\frac{TP}{TP+FN} + \frac{TP}{TP+FP}}{2} \tag{4}$$

The preferred way to summarize sensitivity and specificity in a global accuracy is the correlation coefficient.

$$CC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN) \cdot (TN + FP) \cdot (TP + FP) \cdot (TN + FN)}} \tag{5}$$

It does however have one big drawback, if even one of the four factors in the numerator is zero then the value of the correlation coefficient will be infinity, so it is not used that much.

Instead of the correlation coefficient most studies evaluating gene predictions [2, 3, 8] use the approximate correlation coefficient (6). It approximates the behavior of the correlation coefficient and can take any value in the interval $[-1 : 1]$. A value of -1 means that the prediction was extremely bad, while a perfect prediction receives an AC value of one.

$$AC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right) - 1 \tag{6}$$

2.4.2 Exon Level

At the exon level, as on the nucleotide level, the primary statistical measurements are sensitivity (7) and specificity (8).

$$Sn = \frac{True\ Exons}{Annotated\ Exons} \tag{7}$$

$$Sp = \frac{True\ Exons}{Predicted\ Exons} \tag{8}$$

Due to the complexity of actually predicting an exon correctly, i.e. both ends of the exon, the sensitivity and specificity will probably be rather low. However, a prediction can still be rather good as even a single nucleotide error will be enough to consider an exon as incorrectly predicted. To get a better estimate of how good a prediction is other categories than correctly predicted exons need to be considered. The four [3] categories of exons are,

Correct Both ends of the exon are correctly predicted.

Partial Only one end of the exon is correctly predicted.

Overlap Non of the ends are correctly predicted but a part of the predicted exon overlaps an annotated exon.

Wrong The predicted exon does not overlap any annotated exon.

Using the last three categories it is easy to construct a number of measurements [3] that can be used to better gauge how good a certain prediction is.

$$PCa = \frac{Partial\ Exons}{Annotated\ Exons} \tag{9}$$

$$PCp = \frac{Partial\ Exons}{Predicted\ Exons} \tag{10}$$

$$OLa = \frac{Overlap\ Exons}{Annotated\ Exons} \tag{11}$$

$$WE = \frac{Wrong\ Exons}{Predicted\ Exons} \tag{12}$$

$$ME = \frac{Missed\ Exons}{Annotated\ Exons} \tag{13}$$

The last measurement (13) is just a variation of the forth category.

2.4.3 Gene Level

The primary statistics are once again sensitivity (14) and specificity (15).

$$Sn = \frac{True\ Genes}{Annotated\ Genes} \tag{14}$$

$$Sp = \frac{True\ Genes}{Predicted\ Genes} \tag{15}$$

It is even more complex to predict a gene than an exon as it involves predicting multiple exons correctly. In order to get a good estimate of how well the programs perform in finding genes the measurements mentioned in the previous section (11)–(13) were adapted to the gene level.

$$OL = \frac{Overlap\ Genes}{Annotated\ Genes} \tag{16}$$

$$WG = \frac{Wrong\ Genes}{Predicted\ Genes} \tag{17}$$

$$MG = \frac{Missed\ Genes}{Annotated\ Genes} \tag{18}$$

3 Methods

3.1 DNA to Gene Prediction

The method used for getting the gene predictions differed slightly depending on whether the program was based on Unveil or not. This difference was due to the fact that GeneID and Genscan both read the entire sequence file in one go.

In the end, the method decided upon was the following

1. INPUT: Chromosome Data
2. If program is based on Unveil then goto 3 (a) else goto 3 (b)
3. (a) Split the input into 10 Mbp pieces with 3 Mbp overlap
   (b) Split the input into 400 kbp pieces with 200 kbp overlap
4. Run program on SweGrid
5. Piece the result together and translate it to GTF format
6. Analyze the result with Geval
7. OUTPUT: Textfile containing the result from Geval

The only point that should need any extra explanation is number 3. The split was performed in such a way to make sure that no Ns, a N is used as a placeholder when the true nucleotide is unknown or masked, were kept. All the Ns were skipt and the non-Ns between two Ns were split in the size specified in 3 (a) or 3 (b).

3.2 The Annotation

The annotation, or the set of genes considered correct, was constructed from two pieces of information.

1. UCSC goldenpath assembly, version hg16 [15]
2. NCBI human RefSeq catalogue, Release 4 [16]

The RefSeq [17] catalogue contains the DNA sequence for all known genes but not their chromosomal positions which is what is needed when checking the predictions. The positions were found by running BLAT locally with RefSeq as bait against the human genome assembly. Only the best match for each RefSeq sequence was kept. However, if more than one match had equally high score all of them were saved.

3.3 EST information

The EST sequences [18] were downloaded from NCBI. University of California Santa Cruz (UCSC) has already aligned the ESTs to the human genome so the positions were downloaded from their FTP site [19]. The first experiments were performed with the raw positions that were downloaded but later these positions were postprocessed to close all gaps with a size of 20 bp or less. This was done as only a small portion ($< 0.01\%$) of the human introns are less than 20 bp [20].

3.4 Tweaking Unveil + EST

Although Unveil + EST is working well the numbers used in the EST code, located in *FastViterbi.C*, is not optimal as those were decided upon by testing the code on a small subset of genes. The code in question is shown in the following list.

```
202   if(esttackning<2)
203   {
204       newP+=newP*(0.01);
205       bestP=newP;
206       bestPredecessor=precedingPair.state;
207       continue;
208   }
209   if(esttackning>5)
210   {
211       if(currentPair.state==119)
212           newP+=newP*(-0.0002);
213       else
214           newP+=newP*(-0.0002);
215   }
```

There are four constants that need to be tweaked to optimize the result.

1. The **2** on line 202
2. The **0.01** on line 204
3. The **5** on line 209
4. The **-0.0002** on line 212 and 214

The tweaking was done by testing a lower and a higher number in order to calculate a numerical gradient to see which way the numbers should be changed. The sequence used during the test was the masked copy chromosome 22. Unveil + EST was configured to read 400 kbp segments at a time.

3.5 Analyzing the Result

As no program offered all the things needed for this study it was evident that a new program had to be written. Geval was written from scratch in C++ and used for analyzing the predictions. More information about the program can be found in the Appendix on page on page 26.

3.6 Combining Predictions

The predictions were combined by appending one program's prediction to the end of another program's prediction. This combined prediction was then run through the "Compare Sub-Groups" part of Geval with the search criteria set to "annotation: missed(1)". The number of genes that matched the criteria was recorded. The procedure was then repeated so that all 15 combinations were tested.

3.7 Subgroups: Missed and Found

The prediction for chromosome 22 from Unveil + EST were run through the "Compare Sub-Groups" part of Geval in order to try and figure out why certain genes were discovered while others were missed. The search criteria was set to "annotation:missed(1)".

The two files *true.csv* and *false.csv*, which contains information about the genes that matched the criteria and the genes that did not respectively, was imported into Matlab 6.5 via the import function. The data was then normalized by dividing all the features with the highest number for each feature. An extra toolbox called PRTools [21] was downloaded and installed. To see which of the seven features that separate the two groups best the following command was used:

```
> [I, F] = featrank(A)
```

where A is the dataset containing the information for all the genes.

The seven features considered are,

1. Gene Length (from Start- to Stop-codon)
2. Actual Coding Length
3. Number of Exons
4. Score
5. GC Content
6. Distance to closest gene (Before)
7. Distance to closest gene (After)

4 Results and Discussion

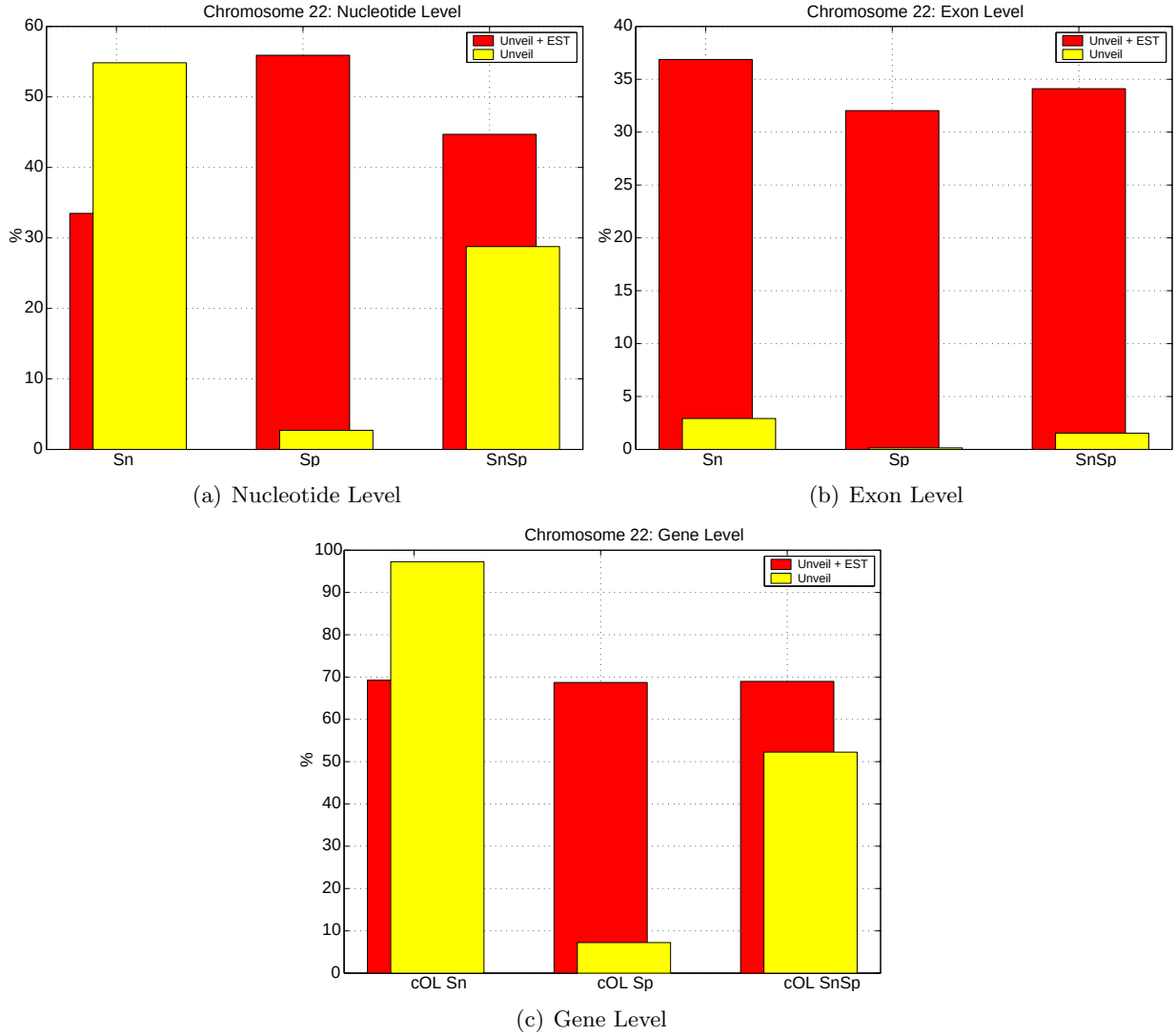
4.1 Comparing Unveil to Unveil + EST

The question whether adding the EST information to Unveil has increased its accuracy or not is as easy to answer as looking at how many genes Unveil predicted before and after adding the information. The original Unveil predicted over 33 thousand genes on chromosome 22 while the new Unveil only found 876. According to the annotation the number of genes is 656. This extreme problem of overpredicting leads to a high number of the coding nucleotides being found (Sn in figure 3(a)). Unveil + EST on the other hand finds only around half as many but close to 60% of its predicted nucleotides are correct compared to only a very small fraction for Unveil (Sp in figure 3(a)).

The improvement is even more visible on the exon level. Unveil seems to have problems when it comes to finding exons as evident by Sn in figure 3(b). It is exactly this problem the EST information is supposed to solve. Unveil + EST uses the ESTs to identify the start and stop of exons and the improvement is easy to see when looking at the exon sensitivity. By using the ESTs

a larger part of the annotated exons are found and also a bigger portion of the predicted exons are true (Sp in figure 3(b)).

Keeping in mind that Unveil predicted over 33 thousand genes it is not hard to realize that the probability for at least one coding nucleotide in each gene being found is high. This leads to a very high cOL Sn (cOL = coding sequence overlap) as seen in figure 3(c). Unveil + EST is not that far behind with close to 70%. However, it should be evident that the performance of Unveil + EST is a lot better than that of Unveil even though it finds almost all the genes. The reason for this is that the genes found by Unveil are hidden among a lot of completely incorrect genes (Sp in figure 3(c)) while around 70% of the genes found by Unveil + EST overlaps an annotated gene.



(a) Nucleotide Level

(b) Exon Level

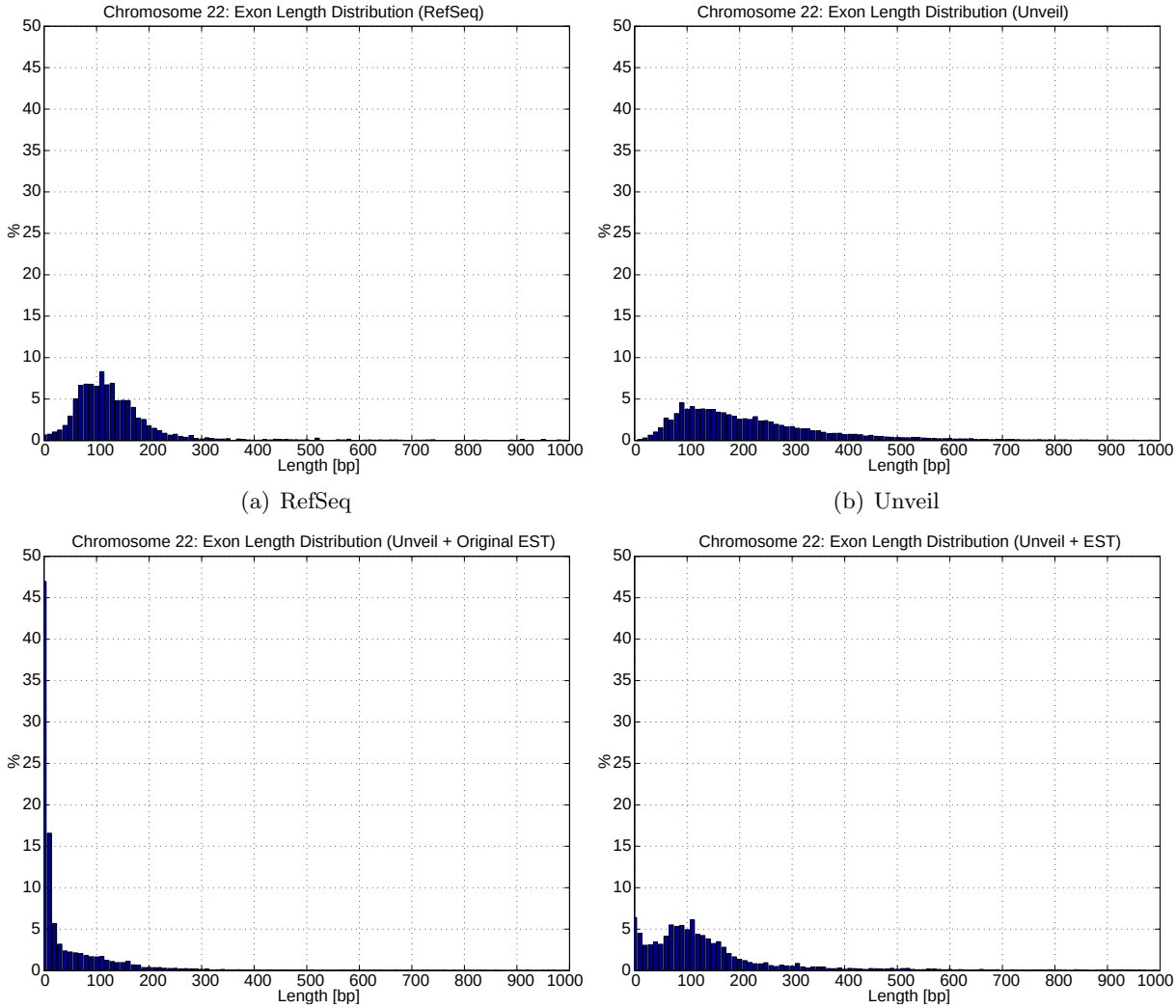
(c) Gene Level

Figure 3: Chromosome 22 - Comparison between Unveil and Unveil + EST

4.2 Internal Exon Length

An important part of predicting genes are the exon length distribution as an incorrect distribution leads to a very low number of correct exons. The length distribution for internal exons should have a wide peak between 100–170 bp and a mean value of around 140 bp [22].

The annotation covering chromosome 22 has a mean value of 165 bp and its distribution can be seen in figure 4(a). None of the different incarnations of Unveil has a perfect exon distribution. The standard Unveil tends, according to figure 4(b), to predict longer exons than it should. This is even more evident when looking at the mean length of the predicted internal exons which is 239 bp almost 100 bp to high. The first attempt at using EST information to improve the result was not a success. Its distribution, figure 4(c), does not look at all like the rest. Instead of having a wide peak it looks almost like an exponential curve. This can be explained by the fact that no postprocessing of the ESTs was performed resulting in lots of potential exons and very short introns separating them. This caused Unveil to shift back and forth between exon and intron states when predicting as it found a start of an exon and then only a few basepairs later found a stop. After postprocessing the ESTs, i.e. joining all exons separated by an intron shorter than 20 bp, the result improved dramatically as seen in figure 4(d). However, the result is still not perfect as Unveil still predicts too many short exons. These short exons are caused by the use of the EST information as they are not visible in the standard Unveil.



(a) RefSeq

(b) Unveil

(c) Unveil + Original EST

(d) Unveil + EST

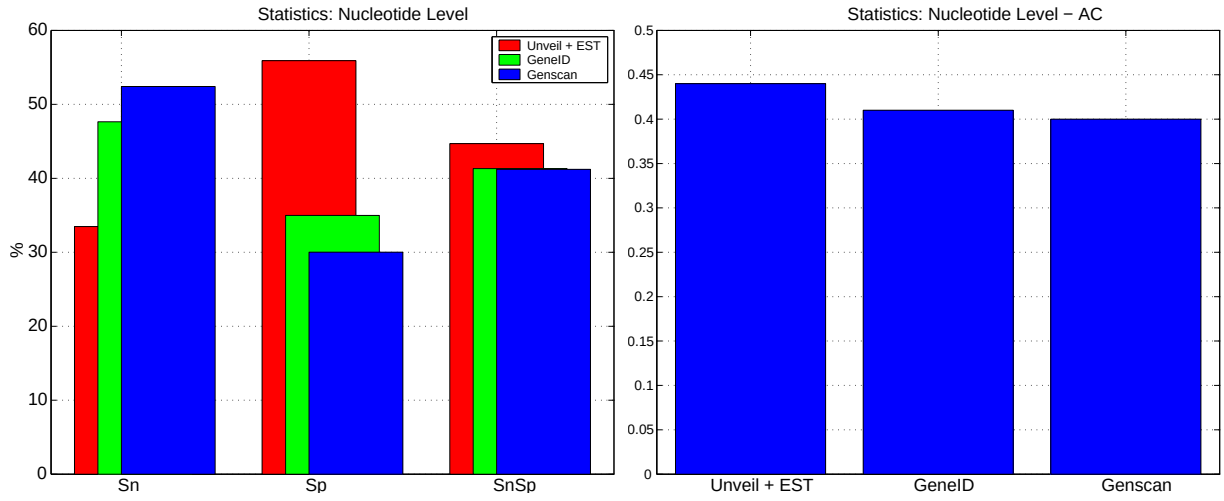
Figure 4: Internal Exon Length Distributions

4.3 Statistical Measurements

Only statistical measurements for Unveil + EST, GeneID, and Genscan are presented as the original Unveil was only run on chromosome 20–22. The reason for this is that Unveil has severe problems with overpredicting making it hard and time consuming to analyze the results.

4.3.1 Nucleotide Level

All three programs performed, according to the approximate coefficient (formula 6), as seen in figure 5(b), equally well. The programs are however good at different things. GeneID and Genscan both identified a larger portion of the coding nucleotides (Sn in figure 5(a)) than Unveil + EST while a smaller part of their predicted nucleotides were correct (Sp in figure 5(b)). This difference is due to GeneID and Genscan predicting more genes than Unveil + EST and thus a higher number of coding nucleotide.



(a) Nucleotide Level

(b) Approximate Coefficient

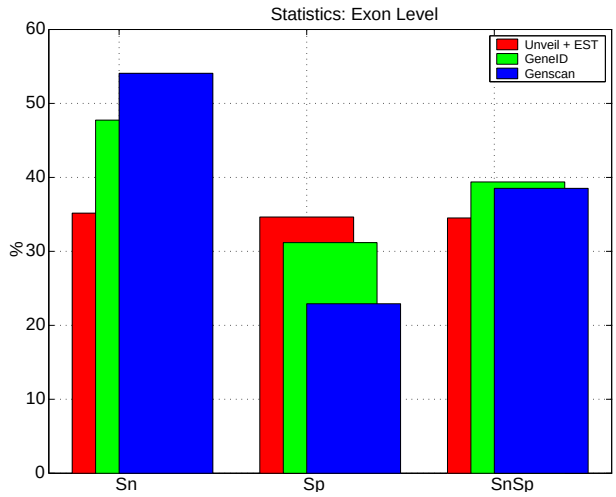
Figure 5: Statistics: Nucleotide Level

4.3.2 Exon Level

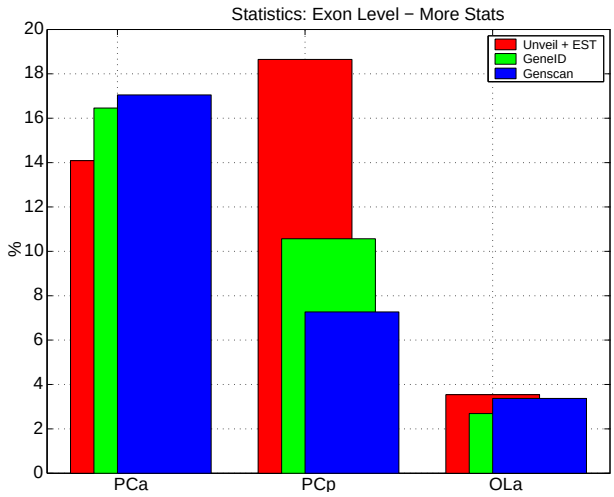
Even though all programs performed equally well on the nucleotide level the same can not be said about the performance on the exon level. Genscan found the highest portion of the annotated exons, (Sn in figure 6(a)) followed closely by GeneID while Unveil + EST was 15% behind the other two. As on the nucleotide level a larger part of Unveil + EST's predictions was true. However, GeneID and Genscan performed almost as well. (Sp in figure 6(a)).

As mentioned, Unveil + EST found 15% fewer annotated exons than the other programs. If you consider the fact that it seemed to predict shorter internal exons than normal it could be expected that some of the exons that were correctly predicted by the other programs ended up as either partial or overlapping exons for Unveil + EST. However, that does not seem to be the case. Looking at figure 6(b) you can clearly see that both GeneID and Genscan predicted more partial exons (PCa) than Unveil + EST and that the difference when it comes to overlapping (OLa) is negligible. The only noticeable difference in figure 6(b) is that a larger part of Unveil + EST predicted exons are partial.

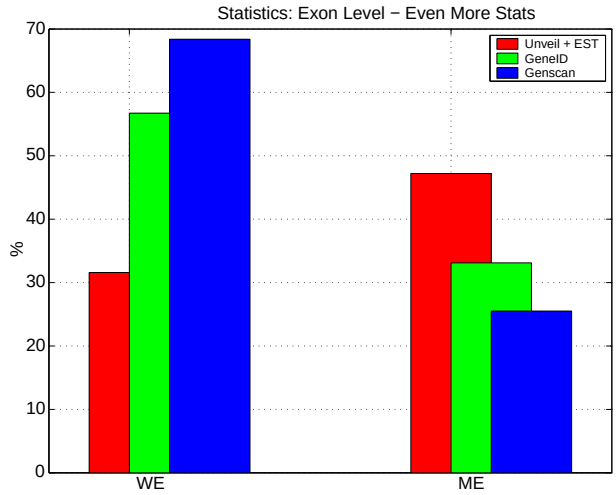
The results when it comes to missing and wrong exons, figure 6(c), can be explained by the fact that GeneID and Genscan overpredict the number of genes. More genes lead to more exons which in turn lead to more wrong and less missed exons. It is important to remember that the annotation is not perfect therefore a missed or wrong exon might actually be an error in the annotation and not an error in the prediction.



(a) Exon Level



(b) Exon Level: More Stats

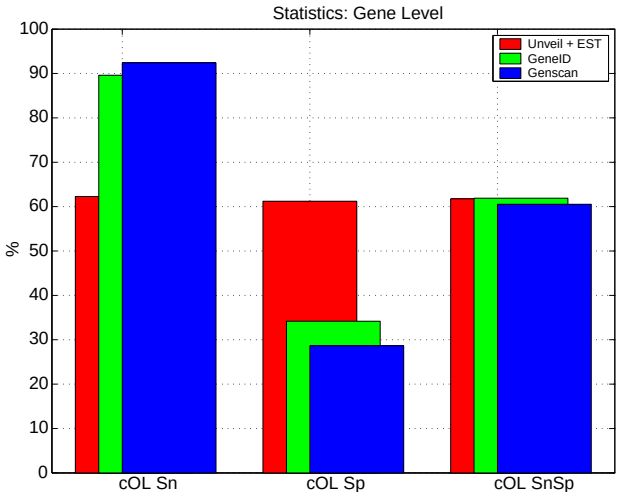


(c) Exon Level: Even More Stats

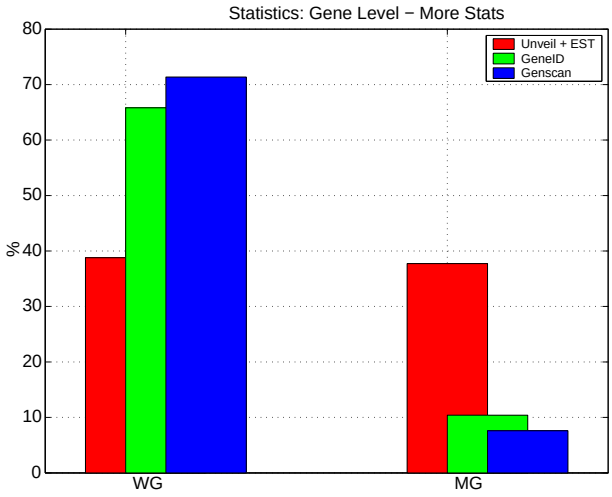
Figure 6: Statistics: Exon Level

4.3.3 Gene Level

Sensitivity and specificity values on the gene level is not presented as no program managed to predict a gene correctly resulting in both Sn and Sp being zero. Instead, coding sequence overlap (cOL) are used to judge the performance on the gene level. GeneID and Genscan finds around 30% more genes than Unveil + EST (cOL Sn in figure 7(a)). But on the other hand, a larger part of Unveil + EST's predictions overlap an annotated gene (cOL Sp in figure 7(a)).



(a) Gene Level



(b) Gene Level: More Stats

Figure 7: Statistics: Gene Level

4.3.4 Splice Sites

No difference between finding 5' or 3' splice sites was detected. As usual, Genscan was the best followed closely by GeneID, with Unveil + EST around 10% behind GeneID (figure 8). In the case of Unveil + EST most of the 5' splice sites that were found are located on the plus strand while a large part of the 3' splice sites are found on the minus strand. This can be explained by remembering the less than perfect exon length distribution (see 4.2) that Unveil + EST had. The 5' splice site is the first thing that Unveil + EST locates when looking for exons on the plus strand. It then fails to find the 3' splice site as the exon length is incorrectly predicted. The same things happens with genes on the minus strand but then the 3' splice site is located first and next the 5' site.

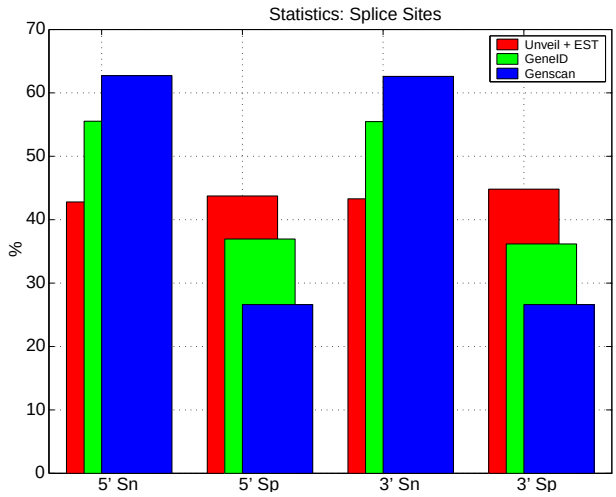


Figure 8: Statistics: Splice Sites

4.4 Combining Predictions

It is quite clear that Unveil + EST is not yet on the same level as GeneID and Genscan. However, if it finds genes that the other two programs missed then its existence is justified. The predictions from all four programs were combined and the number of missed genes are presented in figure 9. The result can in no way be used to judge how good a certain program is as if you predict genes all over the chromosome then the number of missed genes would be zero but the result would be really bad.

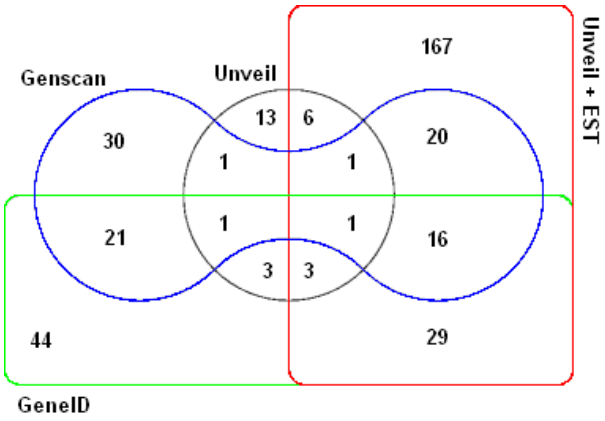


Figure 9: Combining Predictions: Chromosome 22 - Missed Genes

Unveil missed the fewest number of genes. This was expected as it predicted over 33 times the number of genes of the other programs. GeneID and Genscan missed around 30–40 genes while Unveil + EST missed 167 genes.

The interesting results are from the combination of the different predictions. As seen in figure 9, adding the EST information to Unveil caused Unveil to find 7 genes that were not found without the ESTs. Another interesting result is that the combination of Unveil + EST with either GeneID or Genscan, or both, decreased the number of missed genes. This can be taken as evidence that Unveil + EST found genes not predicted by GeneID or Genscan.

One gene is missed even when combining all predictions. This gene was identified as XM_373942.1 (NCBI Identification). Further investigation revealed that the gene was found by a gene prediction program and has later been removed from RefSeq and thus is no longer considered to be a gene.

4.5 Tweaking Unveil + EST

Each constant is responsible for either triggering a change in the probability for an exon or for the actual change.

Unveil + EST decreases the probability for an exon if it does not find enough ESTs that support the exon. How many ESTs that are needed to stop Unveil + EST from decreasing the probability is decided by constant one. It is quite hard to interpret what the numbers, see table 1, mean as they say different things, e.g. the AC value says that decreasing the value of constant one is a good idea while Exon SnSp says the opposite. It all boils down to what you prefer. If you want to find a large portion of the annotated genes then you should lower the constant. On the other hand, if you prefer that the predictions made have a large percentage of true exons then the constant should be increased.

Value	AC	Exon Sn	Exon Sp	Exon SnSp
1	0.51	40.50%	25.32%	32.63%
2	0.48	36.35%	31.89%	33.60%
3	0.48	34.44%	38.25%	35.86%

Table 1: Optimizing Unveil + EST: Constant One

The second constant concerns how large the decrease in probability that an exon exists should be if too few ESTs are found that support the exon. Increasing this constant should therefore result in fewer predicted exons and genes, and the opposite should happen when decreased. This is also what happens, when the constant is at its lowest value the number of genes predicted is 1147, increasing the value to the standard value decreases the number of genes to 1001, and with the highest value only 723 genes are predicted.

It is quite clear from looking at the numbers, see table 2, that the constant should be increased. Even though Unveil + EST predicted fewer genes with the higher value its predictions were of better quality and it even found more exons than before.

Value	AC	Exon Sn	Exon Sp	Exon SnSp
0.0050	0.45	37.20%	30.53%	33.32%
0.0100	0.48	36.35%	31.89%	33.60%
0.0150	0.53	38.49%	35.19%	36.39%

Table 2: Optimizing Unveil + EST: Constant Two

Changing the third constant alters how many ESTs are needed before increasing the probability that an exon exists. An important thing to remember is that the needed number of ESTs are one higher than the value of the constant, so when the value is five the required number of ESTs are six. The numbers, see table 3, shows that the current value probably is the best.

Value	AC	Exon Sn	Exon Sp	Exon SnSp
4	0.48	36.57%	29.78%	32.75%
5	0.48	36.35%	31.89%	33.60%
6	0.45	21.60%	37.96%	29.40%

Table 3: Optimizing Unveil + EST: Constant Three

The forth constant is almost the opposite of constant one, i.e. how much should the probability of an exon increase if it has sufficient EST support. A larger negative value implies a higher increase in the probability than a smaller negative value. The numbers, see table 4, are not as clear this time as before. It looks like the standard number is really bad though. It did not matter in which direction the constant was changed the accuracy increased in both cases. However, it seems that decreasing the constant gives the largest increase in accuracy therefore the constant should probably be decreased.

Value	AC	Exon Sn	Exon Sp	Exon SnSp
-0.00025	0.51	38.15%	34.41%	35.84%
-0.00020	0.48	36.35%	31.89%	33.60%
-0.00015	0.52	37.82%	34.00%	35.48%

Table 4: Optimizing Unveil + EST: Constant Four

All these small tests tell us that the constants should be changed in the following way,

1. No change (or maybe Increase or Decrease)
2. Increase
3. No change
4. Decrease (or maybe Increase)

It is however important to note that the optimal value for each constant probably is dependent on a number of factors that changes depending on what input sequence you want to check and what sequence length Unveil + EST reads. Another important thing is that the four constants are not independent. Constant two and four depend on constant one and three respectively. It is therefore a lot more difficult to optimize the constants than this little test seems to indicate.

4.6 Subgroups: Missed and Found

The text output from the run can be found in the Appendix on page 41. As Unveil + EST does not assign a score to its predictions only six features were considered when trying to figure out why certain genes were missed.

The exact order of how important the different features are for finding the genes varied depending on what technique, or mapping, was used to separate the two groups. But the same features were always near the top regardless of the mapping therefore the standard mapping, which is a neuron net, was used. This resulted in the following order.

1. Number of Exons
2. GC Content
3. Distance to closest gene (Before)
4. Gene Length (from Start- to Stop-codon)
5. Distance to closest gene (After)
6. Actual Coding Length

Looking at the text output it is evident that the number of exons per gene clearly is different between the genes that were found and those that were not (10.28 vs 5.63). It is also easy to draw the conclusion that distance to closest gene (both Before and After) should be higher up in the list as the difference between the genes that were missed and those that were found seems to be

significant (Before: 124k vs 57k, After: 107k vs 61k). However, a good portion of the genes that were not found is clustered together resulting in their distance to closest gene being zero. The same is true for the genes that were found so it is quite hard to use the distance to separate the two groups.

5 Conclusion

Previous studies found that most programs are quite good at predicting nucleotides and exon structure. However, these studies used gene sets that removed most pitfalls that exist when predicting genes using chromosomal data. Things like, long stretches of non-coding DNA, more than one gene in a sequence, genes on both minus and plus strand at the same time, and so on. A good example that programs perform differently when presented with these obstacles is the results from running Unveil. According to a study [10] by the program's authors it predicted 74% of all the exons correctly when tested on 400 *Arabidopsis thaliana* cDNA sequences. In this study Unveil performed the worst, predicting only a very small fraction of all the exons correctly. The older studies are not worthless though, they are good for what they try to do namely comparing different programs to each other but they can not be used for evaluating how well the programs would perform on real data.

The tests performed in this study have shown that the programs are quite good at finding the genes, but that they are far from perfect. GeneID and Genscan found almost all the genes but not a single one was predicted perfectly. Another negative with these programs was that they overpredicted by quite a lot. Unveil + EST on the other hand lagged behind the more mature programs.

Even though Unveil + EST is not as good as GeneID and Genscan it is still a extremely good improvement over the unmodified Unveil. This shows that the way the EST information was implemented in Unveil by Mattias Bäck [13] although it still has some problems, seems to be working quite well. However, the constants that decide how much to increase or decrease the probability that an exon exists are not optimal as shown by the tweaking test. A small change in one of the constants lead to an increase in performance by a few percentage points as shown in table 2 on page 20.

References

[1] The Internationl Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[2] M. Burset and R. Guigo, "Evaluation of Gene Structure Prediction Programs," *Genomics*, vol. 34, pp. 353–367, 1996.

[3] S. Rogic, A. K. Mackworth, and F. B. Ouellette, "Evaluation of Gene–Finding Programs on Mammalian Sequences," *Genome Research*, vol. 11, pp. 817–832, 2001.

[4] R. Guigo, P. Agarwal, J. F. Abril, M. Burset, and J. W. Fickett, "An Assessment of Gene Prediction Accuracy in Large DNA Sequences," *Genome Research*, vol. 10, pp. 1631–1642, 2000.

[5] C. Mathe, M.-F. Sagot, T. Schiex, and P. Rouze, "Current mehtods of gene prediction, their strengths and weaknesses," *Nucleic Aicds Research*, vol. 30, no. 19, pp. 4103–4117, 2002.

[6] J. Frickett and C.-S. Tung, "Assessment of protein coding measures," *Nucleic Acids Research*, vol. 20, pp. 6441–6450, 1992.

[7] J.-M. Claverie, "Computational methods for the identificiation of genes in vertebrate genomic sequences," *Human Molecular Genetics*, vol. 6, no. 10, pp. 1735–1744, 1997.

[8] C. B. Burge and S. Karlin, "Prediction of Complete Gene Structures in Human Genomic DNA," *Journal of Molecular Biology*, vol. 268, pp. 78–94, 1997.

[9] W. H. Majoros, "Unveil: An HMM–based Genefinder for Eukaryotic DNA." http://www.tigr.org/software/Unveil/unveil.pdf (June 13th 2004).

[10] W. H. Majoros, M. Pertea, C. Antonescu, and S. L. Salzberg, "GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic genefinders," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3601–3604, 2003.

[11] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding Genes in DNA with a Hidden Markov Model," *Journal of Computational Biology*, vol. 4, no. 2, pp. 127–141, 1997.

[12] The Sanger Institute, *GFF: an Exchange Format for Feature Description.* http://www.sanger.ac.uk/Software/formats/GFF/ (June 2004).

[13] M. Bäck, "Unknown Title." UNPUBLISHED.

[14] G. Parra, E. Bianco, and R. Guigo, "GeneID in *Drosophila*," *Genome Research*, vol. 10, pp. 511–515, 2000.

[15] UCSC, "UCSC goldenpath assembly, version 16," 2004–04–05. ftp://hgdownload.cse.ucsc.edu/apache/htdocs/goldenPath/hg16/bigZips/chromFa.zip (June 2004).

[16] NCBI, "NCBI human RefSeq catalogue," 2004–04–05. ftp://ftp.ncbi.nih.gov/RefSeq/ (May 2004).

[17] K. D. Pruitt, K. S. Katz, H. Sicotte, and D. R. Maglott, "Introducing RefSeq and LocusLink: curated human genome resources at the NCBI," *Trends in Genetics*, vol. 16, no. 1, pp. 44–47, 2000.

[18] NCBI, "NCBI dbEST version 8," April 2004. ftp://ftp.ncbi.nih.gov/blast/db/FASTA/est_human.gz (May 2004).

[19] UCSC, "BLAT alignments for dbEST." ftp://hgdownload.cse.ucsc.edu/apache/htdocs/ goldenPath/hg16/database/all_est.txt.gz (May 2004).

[20] M. K. Sakharkar, V. T. Chow, and P. Kangueane, "Distributions of exons and introns in the human genome," *In Silico Biology*, vol. 4, no. 32, pp. 1–4, 2004.

[21] Delft University of Technology, "PRTools," 2004–11-01. http://www.prtools.org (Oct 2nd 2004).

[22] J. D. Hawkins, "A survey on intron and exon lengths," *Nucleic Acids Research*, vol. 16, no. 21, pp. 9893–9905, 1988.

6 Acknowledgements

I wish to thank Dr Helgi Schiöth for the opportunity to work on this project. I am also very grateful for the input and help given by Dr Robert Fredriksson, Christian Murray, and Tobias Hill. But most of all I wish to thank Thomas Larsson for all the help he gave me on this project, without his help I do not know how long it would have taken.

Last but not least I wish to thank my room mates Helena Kristiansson and Henry Flisell for all the fun we had and for the fact that they did not strangle me for disturbing them all the time.

7 Appendix

7.1 Program Listings

The following programs have been written during the length of the project. A short description is available for each program.

sweMain.java Splits the sequence files into pieces which length and overlap is specified by the user. It also creates the XRSL files and scripts needed for SweGrid. Written in Java.

MvOutput.java Moves the output received from SweGrid from each individual job directory to one user defined directory. Written in Java.

out2gtf Used to piece together the result from SweGrid and translate the result to the GTF format. Written in C++.

Geval Analyzes the predictions and calculates all the statistics. Written in C++. A longer description can be found in 7.2.

7.2 Geval

7.2.1 Program Description

The reason Geval was written is that the programs investigated did either not offer the features needed or calculated the statistics incorrectly, e.g. more than 100% correct exons during testing.

Geval was written in C++ from scratch and each time the program is run the following algorithm is used. It keeps one list for annotated features and one for predicted.

1. Check that files exist
2. Read FASTA file to find the sequence length
3. Read genes from the annotation
   (a) Read exon, save gene name
   (b) Try to add the exon to the list of unique exons
   (c) If gene name is different than the last exon then
      i. Save last transcript
      ii. Add new splice sites
      iii. Add the transcript's exons to the correct exon type list
      iv. Create a new transcript
      v. Add transcript to gene which matches the current gene name

      else add exon to current transcript
   (d) If end of file has not been reached goto (a).
4. Read genes from the prediction and follow the same method as outlined above for the annotation

5. Compare the unique exons from the annotation with the unique exons from the prediction
6. Compare the annotated transcripts with the predicted transcripts
7. Compare the splice sites
8. Loop over the nucleotides and check the status of each nucleotide in the annotation and its status in the prediction. Add up TN, TP, FN, and FP.
9. Calculate the most time expensive statistics
10. Print all the statistical measurements

The definition of an exact exon is looser than expected. This is due to the inexact nature of BLAT and similar tools. It is not probable that the alignments will be perfect when the annotated genes are BLATed against the human genome. This imperfection makes it hard to believe that the positions are 100% correct. They might be shifted a few bases up or downstream of the actual position. In an attempt to try and rectify this problem an exon is considered exact if its splice sites is almost, within 3 bp, correct. The number 3 bp was selected without much thought.

The same situation is valid for the transcripts. A transcript is considered exact if its exons, using the criteria mentioned in the previous paragraph, matches the exons of an annotated transcript and if its coding length is the same as the annotated transcript. The second condition is important as a shift in position will not result in a shorter transcript.

7.2.2 Classes

The class structure of Geval, see figure 10, is rather complex. A line is drawn from one class to a lower class if the first class includes the other class.

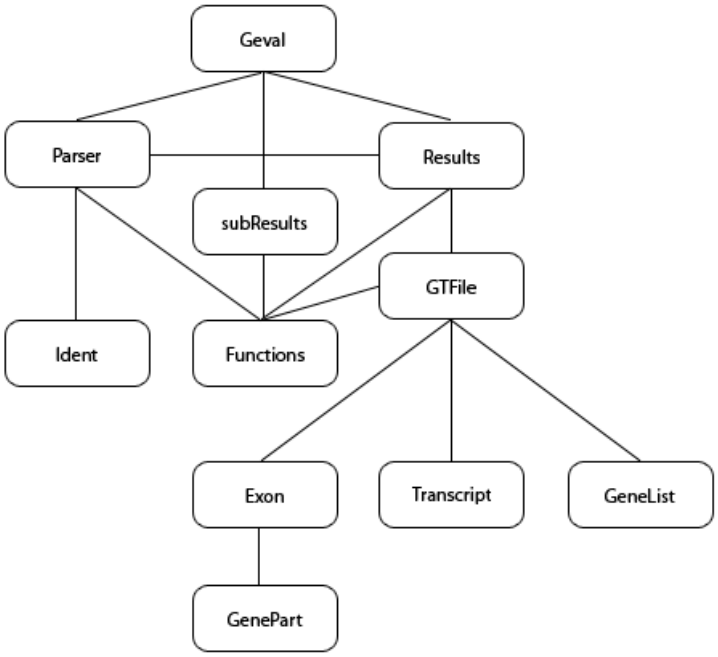


Figure 10: Class diagram for Geval

7.2.3 Statistical Measurements

There are four distinct groups that Geval prints statistics about: transcripts, exons, nucleotides, and signals. A short description about each statistical measurement that Geval outputs for each group is presented in the following lists. An example of the result can be seen in 7.2.4.

Transcripts

Transcripts are divided into three groups, one containing all transcripts one for complete, i.e. both start and stop codon found, and one for incomplete.

Total Length Length is defined as the difference between the 3' end of the last exon and the 5' end of the first exon. Total length is all the lengths added together.

Total Coding Length Total coding length is the length of all exons from all the transcripts added together.

Total Score The sum of the score that the different transcripts got when predicted by the program.

Total Exons The number of exons that the transcripts contain.

Exact A transcript is considered exact if the criteria described in 7.2.1 is fulfilled.

Genomic Overlap During a check for genomic overlap the two transcripts are considered to be one long exon instead of a combination of exons and introns. Then these two long exons are compared to see if they share at least one nucleotide. Exact and genomic overlap is two disjoint groups, i.e. a transcript can not be part of both groups.

CDS Overlap Coding sequence overlap is almost the same as genomic overlap. The only difference is that the actual exons of both transcripts are compared to each other. CDS Overlap is always equal to or smaller than Genomic Overlap.

Wrong A predicted transcript that does not have a single nucleotide in common with an annotated transcript.

Missed An annotated transcript that have no nucleotide in common with a predicted transcript.

Exact Exon Transcripts that contain at least one exact exon.

Start Codon Predicted/Annotated transcripts in this group have their start codon in common with an annotated/predicted transcript.

Stop Codon The same as start codon but in this case the stop codon is considered.

Start Stop Codon A predicted/annotated transcript that matches both the start and stop codon of an annotated/predicted transcript.

Exons

Only unique exons are counted. These are divided into four disjoint groups.

1. **Initial** Exons that contain a start codon but no stop codon
2. **Terminal** Exons that contain a stop codon but no start codon
3. **Single** Exons that contain both a start and stop codon
4. **Internal** The rest

Correct An exon is considered exact is the criteria in 7.2.1 is fulfilled.

Partial A predicted/annotated exon that has one splice site in common with a annotated/predicted exon. Correct and partial are disjoint groups.

Overlap A predicted/annotated exon that has at least one nucleotide in common with an annotated/predicted exon. This group is disjoint with correct and partial.

Overlap 80% A predicted/annotated exon that matches at least 80% of an annotated/predicted exon. The percentage overlap for both exons are calculated as the overlap for the longer exon.

Wrong A predicted exon that does not have a single nucleotide in common with an annotated exon.

Missed An annotated exon that does not have a single nucleotide in common with a predicted exon.

Splice 5 A predicted/annotated exon that have the same 5' splice site as an annotated/predicted exon.

Splice 3 A predicted/annotated exon that have the same 3' splice site as an annotated/predicted exon.

Nucleotides

Nucleotides can be divided into four disjoint groups, see 2.4.2. Four different statistics are calculated from these groups.

1. **Sensitivity** See formula (1) for the definition
2. **Specificity** See formula (3) for the definition
3. **Correlation Coefficient** See formula (5) for the definition
4. **Approximate Coefficient** See formula (6) for the definition

Sensitivity and specificity values are also presented for all different types of exons. However, these numbers are a little biased as a nucleotide can be counted more than once if more than one exon overlaps it.

Signals

Statistics for four different signals were collected. Only unique signal donors and acceptors are considered.

1. **Signal Donor** The 3' end of an initial or internal exon
2. **Signal Acceptor** The 5' end of a terminal or internal exon
3. **Start Codon** No explanation needed
4. **Stop Codon** No explanation needed

The only statistics presented are sensitivity and specificity.

7.2.4 Output Text - Geval

```
Wed Nov 17 17:02:31 2004

** Summary Stats **

   Annotation: Chr22_refseq_new.gtf
   Prediction: Chr22_EST.gtf
Sequence File: chr22.fa

Transcript Sensitivity              0.00%
Transcript Specificity              0.00%
Transcript SnSp                     0.00%
Exon Sensitivity                    36.87%
Exon Specificity                    32.04%
Exon SnSp                           34.11%
Nucleotide Sensitivity              39.02%
Nucleotide Specificity              55.68%
Nucleotide SnSp                     47.35%

** General Stats ** Predictions:
                                  Chr22_refseq_new.gtf     Chr22_EST.gtf
Gene
     All
          Count                              656.00           876.00
          Total Transcripts                  656.00           876.00
          Transcripts Per                      1.00             1.00

Transcript
     All
          Count                              656.00           876.00
          Average Length                   35087.88         46338.94
          Total Length                   23017651.00      40592911.00
          Average Coding Length             2459.30          1468.53
          Total Coding Length            1613300.00       1286436.00
          Average Score                        0.00             0.00
          Total Score                          0.00             0.00
          Exons Per                            9.10             8.77
          Total Exons                       5969.00          7679.00

     Complete
          Count                              656.00           876.00
          Average Length                   35087.88         46338.94
          Total Length                   23017651.00      40592911.00
          Average Coding Length             2459.30          1468.53
          Total Coding Length            1613300.00       1286436.00
          Average Score                        0.00             0.00
          Total Score                          0.00             0.00
          Exons Per                            9.10             8.77
```

Total Exons	5969.00	7679.00
Incomplete		
Count	0.00	0.00
Average Length	0.00	0.00
Total Length	0.00	0.00
Average Coding Length	0.00	0.00
Total Coding Length	0.00	0.00
Average Score	0.00	0.00
Total Score	0.00	0.00
Exons Per	0.00	0.00
Total Exons	0.00	0.00
Exon		
All		
Count	4936.00	5806.00
Average Length	273.25	176.59
Total Length	1348747.00	1025254.00
Average Score	0.00	0.00
Total Score	0.00	0.00
Initial		
Count	562.00	620.00
Average Length	270.36	174.05
Total Length	151943.00	107912.00
Average Score	0.00	0.00
Total Score	0.00	0.00
Internal		
Count	3787.00	4403.00
Average Length	165.01	160.65
Total Length	624889.00	707345.00
Average Score	0.00	0.00
Total Score	0.00	0.00
Terminal		
Count	553.00	647.00
Average Length	889.68	238.64
Total Length	491991.00	154401.00
Average Score	0.00	0.00
Total Score	0.00	0.00
Single		
Count	34.00	154.00
Average Length	2350.71	377.61
Total Length	79924.00	58152.00
Average Score	0.00	0.00
Total Score	0.00	0.00
Nuc		

```
     All
          Count                          1348747.00        1025254.00
     Initial
          Count                           151943.00         107912.00
     Internal
          Count                           624889.00         707345.00
     Terminal
          Count                           491991.00         154401.00
     Single
          Count                            79924.00          58152.00

Signal
     Splice Acceptor
          Count                             4297.00           4761.00
     Splice Donor
          Count                             4292.00           4771.00
     Start Codon
          Count                              656.00            876.00
     Stop Codon
          Count                              656.00            876.00

** Detailed Stats **

Annotation: chr22.fa
Prediction: Chr22_EST.gtf

Transcript
     All
          Count                              876.00
          Ann Count                          656.00
          Average Length                   46338.94
          Total Length                  40592911.00
          Average Coding Length             1468.53
          Total Coding Length            1286436.00
          Average Score                        0.00
          Total Score                          0.00
          Exons Per                            8.77
          Total Exons                       7679.00
          Exact Pred Count                     0.00
          Exact Ann Count                      0.00
          Exact Specificity                    0.00%
          Exact Sensitivity                    0.00%
          Genomic Overlap Pred Count         633.00
          Genomic Overlap Ann Count          489.00
          Genomic Overlap Specificity         72.26%
          Genomic Overlap Sensitivity         74.54%
          CDS Overlap Pred Count             602.00
          CDS Overlap Ann Count              454.00
          CDS Overlap Specificity             68.72%
          CDS Overlap Sensitivity             69.21%
```

```
      Wrong Count                          243.00
      Wrong Specificity                     27.74%
      Missed Count                         167.00
      Missed Sensitivity                    25.46%
      Exact Exon Pred Count                411.00
      Exact Exon Ann Count                 362.00
      Exact Exon Specificity                46.92%
      Exact Exon Sensitivity                55.18%
      Start Codon Pred Count                 6.00
      Start Codon Ann Count                  5.00
      Start Codon Specificity                0.68%
      Start Codon Sensitivity                0.76%
      Stop Codon Pred Count                  8.00
      Stop Codon Ann Count                   4.00
      Stop Codon Specificity                 0.91%
      Stop Codon Sensitivity                 0.61%
      Start Stop Pred Count                  0.00
      Start Stop Ann Count                   0.00
      Start Stop Sensitivity                 0.00%
      Start Stop Specificity                 0.00%

Complete
      Count                                876.00
      Ann Count                            656.00
      Average Length                     46338.94
      Total Length                    40592911.00
      Average Coding Length               1468.53
      Total Coding Length              1286436.00
      Average Score                          0.00
      Total Score                            0.00
      Exons Per                              8.77
      Total Exons                         7679.00
      Exact Pred Count                       0.00
      Exact Ann Count                        0.00
      Exact Specificity                      0.00%
      Exact Sensitivity                      0.00%
      Genomic Overlap Pred Count           633.00
      Genomic Overlap Ann Count            489.00
      Genomic Overlap Specificity           72.26%
      Genomic Overlap Sensitivity           74.54%
      CDS Overlap Pred Count               602.00
      CDS Overlap Ann Count                454.00
      CDS Overlap Specificity               68.72%
      CDS Overlap Sensitivity               69.21%
      Wrong Count                          243.00
      Wrong Specificity                     27.74%
      Missed Count                         167.00
      Missed Sensitivity                    25.46%
      Exact Exon Pred Count                411.00
      Exact Exon Ann Count                 362.00
```

Exact Exon Specificity	46.92%
Exact Exon Sensitivity	55.18%
Start Codon Pred Count	6.00
Start Codon Ann Count	5.00
Start Codon Specificity	0.68%
Start Codon Sensitivity	0.76%
Stop Codon Pred Count	8.00
Stop Codon Ann Count	4.00
Stop Codon Specificity	0.91%
Stop Codon Sensitivity	0.61%
Start Stop Pred Count	0.00
Start Stop Ann Count	0.00
Start Stop Sensitivity	0.00%
Start Stop Specificity	0.00%

Incomplete

Count	0.00
Ann Count	0.00
Average Length	0.00
Total Length	0.00
Average Coding Length	0.00
Total Coding Length	0.00
Average Score	0.00
Total Score	0.00
Exons Per	0.00
Total Exons	0.00
Exact Pred Count	0.00
Exact Ann Count	0.00
Exact Specificity	0.00%
Exact Sensitivity	0.00%
Genomic Overlap Pred Count	0.00
Genomic Overlap Ann Count	0.00
Genomic Overlap Specificity	0.00%
Genomic Overlap Sensitivity	0.00%
CDS Overlap Pred Count	0.00
CDS Overlap Ann Count	0.00
CDS Overlap Specificity	0.00%
CDS Overlap Sensitivity	0.00%
Wrong Count	0.00
Wrong Specificity	0.00%
Missed Count	0.00
Missed Sensitivity	0.00%
Exact Exon Pred Count	0.00
Exact Exon Ann Count	0.00
Exact Exon Specificity	0.00%
Exact Exon Sensitivity	0.00%
Start Codon Pred Count	0.00
Start Codon Ann Count	0.00
Start Codon Specificity	0.00%
Start Codon Sensitivity	0.00%

```
          Stop Codon Pred Count                    0.00
          Stop Codon Ann Count                     0.00
          Stop Codon Specificity                   0.00%
          Stop Codon Sensitivity                   0.00%
          Start Stop Pred Count                    0.00
          Start Stop Ann Count                     0.00
          Start Stop Sensitivity                   0.00%
          Start Stop Specificity                   0.00%

Exon
     All
          Count                                 5806.00
          Ann Count                             4936.00
          Average Length                         176.59
          Total Length                        1025254.00
          Average Score                             0.00
          Total Score                               0.00
          Correct Pred Count                    1860.00
          Correct Ann Count                     1820.00
          Correct Specificity                     32.04%
          Correct Sensitivity                     36.87%
          Partial Pred Count                    1269.00
          Partial Ann Count                      926.00
          Partial Specificity                     21.86%
          Partial Sensitivity                     18.76%
          Overlap Pred Count                    1119.00
          Overlap Ann Count                      218.00
          Overlap Specificity                     19.27%
          Overlap Sensitivity                      4.42%
          Overlap 80% Pred Count                  31.00
          Overlap 80% Ann Count                   26.00
          Overlap 80% Specificity                  0.53%
          Overlap 80% Sensitivity                  0.53%
          Wrong Exons                           1558.00
          Wrong Exons Specificity                 26.83%
          Missed Exons                          1972.00
          Missed Exons Sensitivity                39.95%
          Splice 5 Pred Count                   2515.00
          Splice 5 Ann Count                    2347.00
          Splice 5 Specificity                    43.32%
          Splice 5 Sensitivity                    47.55%
          Splice 3 Pred Count                   2510.00
          Splice 3 Ann Count                    2340.00
          Splice 3 Specificity                    43.23%
          Splice 3 Sensitivity                    47.41%

     Initial
          Count                                  620.00
          Ann Count                              562.00
          Average Length                         174.05
```

Total Length	107912.00
Average Score	0.00
Total Score	0.00
Correct Pred Count	52.00
Correct Ann Count	29.00
Correct Specificity	8.39%
Correct Sensitivity	5.16%
Partial Pred Count	169.00
Partial Ann Count	175.00
Partial Specificity	27.26%
Partial Sensitivity	31.14%
Overlap Pred Count	120.00
Overlap Ann Count	37.00
Overlap Specificity	19.35%
Overlap Sensitivity	6.58%
Overlap 80% Pred Count	10.00
Overlap 80% Ann Count	7.00
Overlap 80% Specificity	1.61%
Overlap 80% Sensitivity	1.25%
Wrong Exons	279.00
Wrong Exons Specificity	45.00%
Missed Exons	321.00
Missed Exons Sensitivity	57.12%
Splice 5 Pred Count	83.00
Splice 5 Ann Count	35.00
Splice 5 Specificity	13.39%
Splice 5 Sensitivity	6.23%
Splice 3 Pred Count	192.00
Splice 3 Ann Count	200.00
Splice 3 Specificity	30.97%
Splice 3 Sensitivity	35.59%

Internal

Count	4403.00
Ann Count	3787.00
Average Length	160.65
Total Length	707345.00
Average Score	0.00
Total Score	0.00
Correct Pred Count	1749.00
Correct Ann Count	1781.00
Correct Specificity	39.72%
Correct Sensitivity	47.03%
Partial Pred Count	910.00
Partial Ann Count	566.00
Partial Specificity	20.67%
Partial Sensitivity	14.95%
Overlap Pred Count	855.00
Overlap Ann Count	104.00
Overlap Specificity	19.42%

```
     Overlap Sensitivity                    2.75%
     Overlap 80% Pred Count                14.00
     Overlap 80% Ann Count                 13.00
     Overlap 80% Specificity                0.32%
     Overlap 80% Sensitivity                0.34%
     Wrong Exons                          889.00
     Wrong Exons Specificity               20.19%
     Missed Exons                        1336.00
     Missed Exons Sensitivity              35.28%
     Splice 5 Pred Count                 2237.00
     Splice 5 Ann Count                  2132.00
     Splice 5 Specificity                  50.81%
     Splice 5 Sensitivity                  56.30%
     Splice 3 Pred Count                 2205.00
     Splice 3 Ann Count                  2101.00
     Splice 3 Specificity                  50.08%
     Splice 3 Sensitivity                  55.48%

Terminal
     Count                                647.00
     Ann Count                            553.00
     Average Length                       238.64
     Total Length                      154401.00
     Average Score                          0.00
     Total Score                            0.00
     Correct Pred Count                    63.00
     Correct Ann Count                     10.00
     Correct Specificity                    9.74%
     Correct Sensitivity                    1.81%
     Partial Pred Count                   174.00
     Partial Ann Count                    185.00
     Partial Specificity                   26.89%
     Partial Sensitivity                   33.45%
     Overlap Pred Count                   131.00
     Overlap Ann Count                     70.00
     Overlap Specificity                   20.25%
     Overlap Sensitivity                   12.66%
     Overlap 80% Pred Count                 4.00
     Overlap 80% Ann Count                  6.00
     Overlap 80% Specificity                0.62%
     Overlap 80% Sensitivity                1.08%
     Wrong Exons                          279.00
     Wrong Exons Specificity               43.12%
     Missed Exons                         288.00
     Missed Exons Sensitivity              52.08%
     Splice 5 Pred Count                  191.00
     Splice 5 Ann Count                   180.00
     Splice 5 Specificity                  29.52%
     Splice 5 Sensitivity                  32.55%
     Splice 3 Pred Count                  109.00
```

```
          Splice 3 Ann Count                    39.00
          Splice 3 Specificity                  16.85%
          Splice 3 Sensitivity                   7.05%

     Single
          Count                                154.00
          Ann Count                             34.00
          Average Length                       377.61
          Total Length                       58152.00
          Average Score                          0.00
          Total Score                            0.00
          Correct Pred Count                     4.00
          Correct Ann Count                      0.00
          Correct Specificity                    2.60%
          Correct Sensitivity                    0.00%
          Partial Pred Count                    20.00
          Partial Ann Count                      0.00
          Partial Specificity                   12.99%
          Partial Sensitivity                    0.00%
          Overlap Pred Count                    19.00
          Overlap Ann Count                      7.00
          Overlap Specificity                   12.34%
          Overlap Sensitivity                   20.59%
          Overlap 80% Pred Count                 3.00
          Overlap 80% Ann Count                  0.00
          Overlap 80% Specificity                1.95%
          Overlap 80% Sensitivity                0.00%
          Wrong Exons                          111.00
          Wrong Exons Specificity               72.08%
          Missed Exons                          27.00
          Missed Exons Sensitivity              79.41%
          Splice 5 Pred Count                   15.00
          Splice 5 Ann Count                     0.00
          Splice 5 Specificity                   9.74%
          Splice 5 Sensitivity                   0.00%
          Splice 3 Pred Count                   13.00
          Splice 3 Ann Count                     0.00
          Splice 3 Specificity                   8.44%
          Splice 3 Sensitivity                   0.00%

Nuc
     Detailed
          True Positive Count                511161.00
          True Negative Count              97077156.00
          False Positive Count               406948.00
          False Negative Count               798679.00

     Statistics
          Specificity                           55.68%
          Sensitivity                           39.02%
```

```
          SnSp                                47.35%
          Correlation Coefficient             -0.00
          Approximate Coefficient              0.47

     All
          Count                           1025254.00
          Ann Count                       1348747.00
          Correct Pred Count               591102.00
          Correct Ann Count                530986.00
          Correct Specificity                 57.65%
          Correct Sensitivity                 39.37%

     Initial
          Count                            107912.00
          Ann Count                        151943.00
          Correct Pred Count                46701.00
          Correct Ann Count                 48406.00
          Correct Specificity                 43.28%
          Correct Sensitivity                 31.86%

     Internal
          Count                            707345.00
          Ann Count                        624889.00
          Correct Pred Count               480146.00
          Correct Ann Count                353280.00
          Correct Specificity                 67.88%
          Correct Sensitivity                 56.53%

     Terminal
          Count                            154401.00
          Ann Count                        491991.00
          Correct Pred Count                61267.00
          Correct Ann Count                126554.00
          Correct Specificity                 39.68%
          Correct Sensitivity                 25.72%

     Single
          Count                             58152.00
          Ann Count                         79924.00
          Correct Pred Count                 5447.00
          Correct Ann Count                  2746.00
          Correct Specificity                   9.37%
          Correct Sensitivity                   3.44%

Signal
     Signal Acceptor
          Count                              4761.00
          Ann Count                          4297.00
          Correct Pred Count                    0.00
          Correct Ann Count                     0.00
```

```
          Correct Specificity                    0.00%
          Correct Sensitivity                    0.00%

     Signal Donor
          Count                               4771.00
          Ann Count                           4292.00
          Correct Pred Count                     0.00
          Correct Ann Count                      0.00
          Correct Specificity                    0.00%
          Correct Sensitivity                    0.00%

     Start Codon
          Count                                876.00
          Ann Count                            656.00
          Correct Pred Count                     6.00
          Correct Ann Count                      5.00
          Correct Specificity                    0.68%
          Correct Sensitivity                    0.76%

     Stop Codon
          Count                                876.00
          Ann Count                            656.00
          Correct Pred Count                     8.00
          Correct Ann Count                      4.00
          Correct Specificity                    0.91%
          Correct Sensitivity                    0.61%

Wed Nov 17 17:02:31 2004
```

7.3 Output Text - Subgroups: Missed and Found

```
Sub-group Results

** Information **

Criteria: annotation:missed(1)

 Group 1: Annotation (True)
 Group 2: Annotation (False)
 Group 3: Prediction (All)

** General Stats **
                                   Group 1         Group 2         Group 3

     Count                          167.00          489.00          876.00
     Average Length               26225.38        38114.55        46338.94
     Total Length               4379638.00     18638013.00     40592911.00
```

Average Coding Length	2038.14	2603.13	1468.53
Total Coding Length	340370.00	1272930.00	1286436.00
Average Score	0.00	0.00	0.00
Total Score	0.00	0.00	0.00
Exons Per	5.63	10.28	8.77
Total Exons	940.00	5029.00	7679.00

** Detailed Stats **

	Group 1	Group 2	Group 3
Exon			
All			
Count	940.00	5029.00	7679.00
Average Length	362.10	253.12	167.53
Total Length	340370.00	1272930.00	1286436.00
Average Score	0.00	0.00	0.00
Total Score	0.00	0.00	0.00
Initial			
Count	149.00	473.00	722.00
Average Length	282.17	291.60	166.02
Total Length	42043.00	137925.00	119866.00
Average Score	0.00	0.00	0.00
Total Score	0.00	0.00	0.00
Internal			
Count	624.00	4067.00	6081.00
Average Length	186.36	156.09	154.91
Total Length	116291.00	634824.00	942027.00
Average Score	0.00	0.00	0.00
Total Score	0.00	0.00	0.00
Terminal			
Count	149.00	473.00	722.00
Average Length	877.39	996.96	230.46
Total Length	130731.00	471562.00	166391.00
Average Score	0.00	0.00	0.00
Total Score	0.00	0.00	0.00
Single			
Count	18.00	16.00	154.00
Average Length	2850.28	1788.69	377.61
Total Length	51305.00	28619.00	58152.00
Average Score	0.00	0.00	0.00
Total Score	0.00	0.00	0.00
Transcript			
Average GC Content	52.32%	51.68%	51.47%

Average Distance Before	124049.50	57382.15	40253.12
Total Distance Before	20592217.00	28002488.00	35221480.00
Average Distance After	106867.76	60932.74	40567.48
Total Distance After	17846916.00	29796109.00	35537115.00