Finding genes by computational methods:
An analysis of the methods and programs

George Sen
Biochemistry 218
Gsen@stanford.edu
June 6, 2002

Introduction

And the end of all our exploring
will be to arrive where we started
and know the place for the first time.
T.S. Eliot (1942)

True to this statement from T.S. Eliot, we have journeyed to the peak of science by forging a blueprint of the human genome. At the end of this amazing journey, we must begin again where we started. Now, in a different direction to understand what the genes in our genome encode. With nearly 3 billion base pairs, the human genome is modern day molecular wonder. The biggest challenge now is trying to find the genes or coding regions which comprise only 1-3% of the genome in a network of introns comprising 24% and intergenic regions comprising 75% of the human genome(Venter et al). Luckily, there are rules that govern coding regions which can be identified by both experimental and computational means. This report will focus on computational methods used currently to identify genes in the genome.

There are two main methods for computational gene identification which include sequence similarity searches while the other method is gene structure and signal searches also referred to as ab initio gene finding(Rogic et al). Sequence similarity searches is a conceptually simple approach which is based on finding similarity in gene sequences between ESTs(expressed sequence tag), proteins, or other genomes to the input genome. The idea here is that exons are more likely to be conserved due to functional constraints whereas intergenic or intronic regions can rapidly diverge. Once there is similarity by a certain genomic region and an EST, DNA, or protein the similarity information can be used to infer gene structure or function of that region. EST based sequence similarity usually has drawbacks in that ESTs only correspond to small portions of the gene sequence which means that it is often difficult to predict the complete gene structure of a given region(Rogic et al). The biggest downside to this type of approach is that only about half the genes being discovered have significant homology to genes in the databases.

The second method for the computational identification of genes is to use gene structure as a template to detect genes. Coding sequences have a statistical regularity that can be used to our advantage. One example of this is codon bias particularly dicodon counts. Researchers have found that there is a tendency to have the same nucleotide appearing every 3,6,9.. bp in open reading frames(Fickett). Programs such as GeneMark.hmm, Genscan, and Hmmgene use this type of regularity as the basis for gene finding(Rogic et al). Another method is to use signal sensors.These gene finding programs could search for promoter elements, start and stop codons, splice sites, or poly-A sites(Guigo).

In this report, two types of ab initio programs(Genscan and HMMgene) and one program that combines sequence similarity and ab initio(Procruste) will be discussed. A discussion about two separate programs that were used to find genes in the human genome will also be performed.

Genescan

One of the better known and effective ab initio programs was developed by Chris Burge and Samuel Karlin in 1997. This program uses a generalized, fifth-order Markov model in order to discover exons, introns, their splice sites as well as promoter regions. The signals for the sites mentioned above are modeled by weight matrices, weight arrays, and maximal dependence decomposition. Genescan is also different from earlier ab initio programs in that it looks for genes on both strands of DNA simultaneously as well as being able to search sequences for partial, complete, or multiple genes.

Several features of genomic structure were taken into account in Genescan. An example of this is that different length distribution functions were used for the determination of initial, internal, and terminal exons. The distribution of lengths of the different types of exons tend to be different indicating that there are structural constraints that govern efficient splicing. (Burge et al) As mentioned previously, Genescan uses a maximal dependence decomposition procedure to predict splice signals. Previously described programs use a weight matrix method(WMM) or weight array model(WAM) to predict splice signals. The WMM written by Staden et al(1984) assumes no dependence between adjacent as well as non-adjacent donor splice signals whereas the WAM method assumes a dependence only on adjacent positions. The authors of Genescan found there to be significant dependencies between adjacent as well as non-adjacent splice signals which was modeled by the MDD.(Burge et al) Genescan was trained on a set of non-redundent human genomic sequences(2,580,965 bp) with putative, alternatively spliced, viral, and pseudogenes filtered out.

HMMgene

HMMgene uses a standard HMM with coding regions being modeled by 4th order inhomogeneous Markov chains (Krogh 1997). The program is trained using the conditional maximum likelihood criterion which allows for maximizing correct predictions(Rogic et al). There can be several predictions with each block which allows for prediction of alternative splicing. The training set includes human sequences taken from GenBank and put together by Kulp et al. to train Genie. The main difference between this type of HMM ab initio program versus another HMM ab initio program such as Genscan is that HMMgene uses standard HMM while Genscan uses generalized HMM. One advantage of a generalized HMM is that different sensors can be modeled by any type such as neural networks whereas the sensors for a standard HMM program is limited to an HMM framework. On the contrary, an advantage of HMMgene is that it is an integrated model(Burge 1997)(Krogh 1997).

Procruste

Procruste is a program that combines both ab initio as well as sequence similarity searches which was developed by Mikhail Gelfand and Pavel Pevzner in 1996. The strategy of this program is to first eliminate improbable exon containing sites by discarding any sequences that do not contain potential donor and acceptor sites for slicing. This is based on consensus sequences for donor sites which is GU and acceptor sites having a dinucleotide of AG. From the reduced/filtered set of blocks, all combinations of exons are assembled and then compared for similarity to known proteins. The program was able to assemble 87% of the exons correctly in the human genomic test set used where the homologous protein was known. There a couple of weaknesses to this program. One resides in that sequence similarity programs need known homologous proteins to be effective. Since it is estimated that 50% of the proteins in the human genome has no currently known homologs, sequence similarity programs are limited in their usefullness(Rogic 2001). Another potential problem is that the program initially filters out any blocks without the "accepted" consensus donor and acceptor splice sites. This may actually underrepresent the number of true exons since there are other signals that regulate splicing.

Otto

The Otto system was developed by Celera Genomics to annotate the human genome. The program consists of two types of methods to find genes. The first approach is to go through the genome and annotate genes that are high similarity matches to already known human genes. The entire list of currently known human genes has been compiled and is referred to as RefSeq. The cutoff for annotation of a gene when comparing to RefSeq is that the genomic sequnce has to match at least 50% of its length to the RefSeq. The sequence identity must be greater than 92%. They annotated 6528 genes by this method.

The second approach that is employed by this program is to compare the human genomic sequences to EST, protein, and genomic sequence databases. These searches are performed using BLAST. Otto makes a comparison to the EST database by finding matches between the genomic sequence compared to rodent and human ESTs. It also searches protein databases by first translating the human genomic sequences and then making the comparison. The program compares the genomic sequence between mouse and humans to identify potential coding regions that have been conserved. Regions that have homology to any of the criteria above are marked and then two types of analysis performed on them. One type of analysis is to use Genscan to predict the gene structure in the regions marked. The other method is to directly compare each predicted gene to the homology based evidence. For internal regions of first and last exons, there must be homology to within 10 bases while the external regions are allowed more divergence. Internal exons must be supported by homology to within +/- 10 bases of their edges. To evaluate the predictive power of each individual component of the program, they tested the specificity and sensitivity of the RefSeq, homology, and Genscan searches to each other. The specificity is measured by taking the number of correctly matched bases divided by the sum of the number of incorrectly and correctly matched bases while the sensitivity is the number of correctly matched bases divided by the length of the cDNA that the sequence is being compared. When these comparisons are made, the Otto program using RefSeq outscores Otto using a homology based approach, and Genscan. (See Table 1) With a specificity of 0.973 and sensitivity of 0.939 Otto using RefSeq is the ideal program but of course the drawback to RefSeq is that only known genes can be used and thus annotation of novel genes is not possible. The combination of the methods listed above yielded 17,764 anotated genes.

Table 1. Comparison of sensitivity and specificity between Otto(RefSeq and homology) and Genscan

METHOD	SENSITIVITY	SPECIFICTY
Otto(Refseq)	0.939	0.973
Otto(homology)	0.604	0.884
Genscan	0.501	0.633

Reproduced from Venter et al.

Because of the conservative way the genome was annotated, the authors decided to take another approach to potentially identify more genes. They took all the previous regions where there was homology to ESTs, proteins, or mouse genome but did not make the cutoff and ran three de novo gene finding programs. These programs included Grail, Genscan, and FgenesH. They received 155,695 predictions of genes but roughly half was non-redundant, 76,410. Out of the 76,410 only 57,935 did not overlap genes already annotated by Otto or match known genes. Of the 57,935 genes only 21,350 were supported by one type of sequence similarity data and only 8,619 were supported by two or more. The estimated number of genes in the human genome is predicted to be the number of Otto predicted genes,17,764 plus the genes found by the three gene finding programs 21,350 or 8,619. From these estimates, the number of genes in the genome ranges from 26,383 to 39,114(Venter et al).

Exofish

Exofish(**Exo**n **fi**nding by **s**equence **h**omology) is based on sequence homology searches to identify genes in the human genome. It uses the genome of another vertebrate, Teraodon nigroviridis, a type of pufferfish to find sequence similarities to the human genome. The advantage of the pufferfish is in the structure of its genome which is eight times more compact than the human (Crollius et al). Because of this, its intron size ranges from 47bp to 1,476bp while human introns are between 131bp and 12,286bp(Baxendale et al). Since the two organisms are separated by 400 million years of evolution, the essential coding regions should stay conserved whereas the introns would have the flexibility to change. The concept of using a vertebrate with a compact genome to find genes in humans originated in 1995 in a paper by Baxendale et al. They compared the Huntington's disease(HD) gene in pufferfish and human. They found that the pufferfish HD gene only spanned 23kb of genomic DNA whereas the human version spanned 170kb. Despite the large size discrepency, all 67 exons of the HD gene was conserved. All homology searches between the genomes were done using the BLAST algorithm. They found that the best specificty and sensitivity was achieved using TBLASTX where there needed to be a minimum of five consecutive matches and less than two consecutive mismatches in amino acids. To decrease the computation time, the Blosum 62 matrix was replaced by a matrix that just uses two values which include matches and mismatches.The general scheme of Exofish is illustrated in Figure 1.

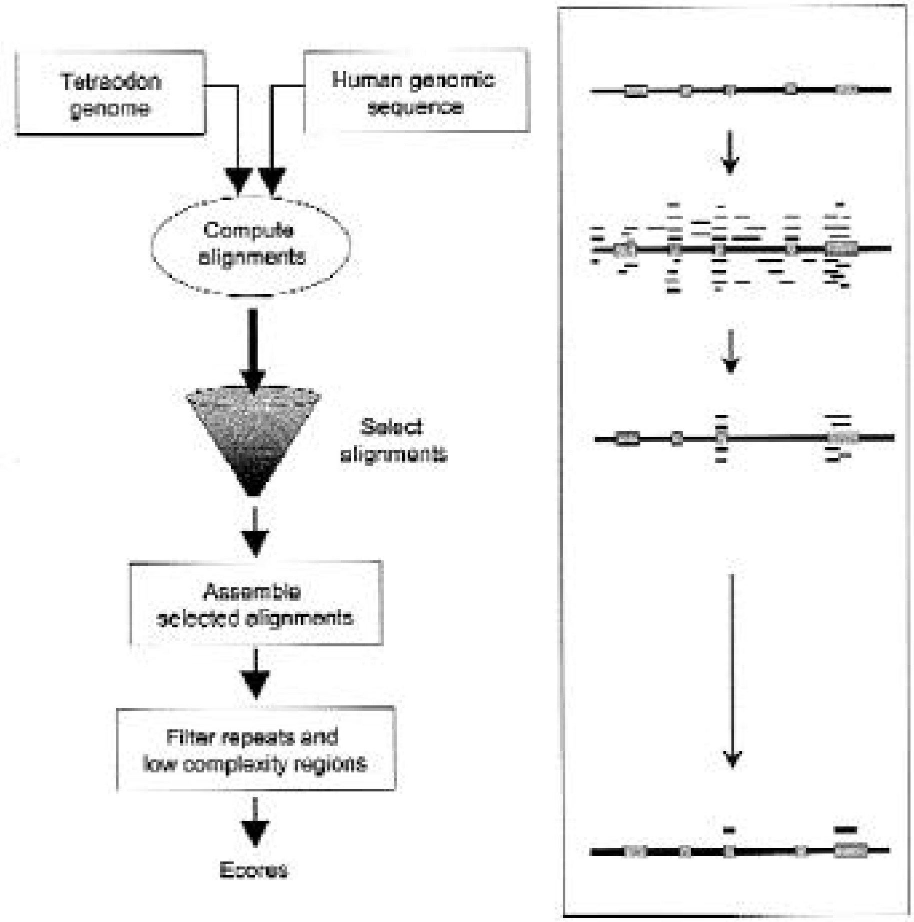

Figure 1. Flow diagram of Exofish
Taken from Crollius et al.

The authors of the paper performed their test of this method by using a set of 4,888 human cDNA sequences taken from Unigene. The Exofish method could detect 70% of the genes with each gene containing about 3.18 ecores which stands for evolutionary conserved region(Crollius et al). Exofish was also used to analyze human chromosome 22. They found 1,525 ecores and 1,344 of those were within annotated regions. Of those regions, 1,197 or 89% were annotated genes and 147 or 11% were within pseudogenes. The next step was to predict the number of genes in the full human genome using Exofish. Since the genome was not completed at the time of the analysis of this method, the authors used the working draft sequence. The working draft had 1,272.3 Mb of human DNA which is about 42.4% of the genome. Exofish found 42,066 ecores and since there are an average of 3.18 ecores per gene, the number of genes in 42.4% of the genome is 11,722 with the consideration that 89% are genes. Thus, the complete genome would have about 27,767 genes. They also set an upper limit on the number of genes by using an ecore value of 2.58 which they found on smaller test sets. This would lead to a lower estimate of 27,767 to an upper limit of 34,224 genes in the genome. These values are very similar to those predicted by Otto. There are definitely weaknesses to this method since everything is based on sequence homology to the pufferfish. One major weakness in this type of approach is that one would not be able to find newly formed or diverged genes between the genomes.

Concluding Remarks

To date there is no perfect gene prediction program. Every program has its own limitations and drawbacks. But, over the recent years, these programs have improved in sensitivity and specificty in predicting gene structure. Many of these programs have improved due to using combinatorial approaches to finding genes such as integrating sequence similarity, signal sensors, and codon bias. Even by using combinatorial approaches the results are not ideal. For example, using Otto(homology) the sensitivity and specificity are not extremely high at 0.604 and 0.884 respectively. Genscan performed even worse at a sensitivity of 0.501 and a specificity of 0.633. These relatively low numbers come from our lack of understanding of all the signals and structural information for genes. Until that is more clearly understood, the accuracy in the predictions will not rise significantly. Another problem with current gene prediction programs is that most do not have the capability to look for alternatively spliced transcripts. The ability to predict that would require much more knowledge on the regulation of gene splicing. One possible method to approach the issue of alternatively spliced transcripts or even finding genes that have no known homologs is by an experimental approach. In order to do this, one would have to use a cDNA subtraction approach to look for tissue specific transcripts. Tissue specific transcripts would most likely account for alternatively spliced genes. Different cell types from different tissues can be isolated and then cDNA libraries made from them while subtracting out the transcripts that are the same in other tissue types. These subtracted libraries can then be sequenced and their coding regions in the genome can be found. There are countless other ways to find genes in the genome and as time passes more methods will be developed to further refine and define the genes that make us who we are today.

References

Baxendale S et al. Comparative sequence analysis of the human and pufferfish Huntington’s disease gene. *Nature Genetics* **10**, 67-75(1995).

Burge C. and Karlin S. Prediction of Complete Gene Structures in Human Genomic DNA. *J Mol Biol* **268**, 78-94(1997).

Crollius HR et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics* **25**, 235-238(2000).

Fickett JW. Finding genes by computer: the state of the art. *TIG* **12,** 316-320(1996).

Gelfand M, Mironov A, and Pevzner P. Gene recognition via spliced sequence alignment. *Proc Natl Acad Sci USA* **93,** 9061-9066(1996).

Guigo R. Computational Gene Identification. *J Mol Med* **75,** 389-393(1997).

Guigo R. et al. An Assessment of Gene Prediction Accuracy in Large DNA Sequences. *Genome Research*, 1631-1642(2000).

Krogh A. Gene finding: putting the parts together. Guide to Human Genome Computing. 261-274(1998).

Rogic S, Mackworth A, and Ouellette F. Evaluation of Gene-Finding Programs on Mammalian Sequences. *Genome Research*, 817-832(2001).

Staden, R. Computer Methods to Locate Signals in Nucleic Acid Sequences. *Nucl Acids Res* **12**, 505-519(1984).

Venter C. et al. The Sequence of the Human Genome. *Science* **291**, 1304-1351(2001).