Gene identification is still extremely difficult for many organisms. The number of genes for any given genome is not fixed in stone, as alternative splice sites and regulatory regions come to light.

Genome projects, and smaller sequencing studies, use gene identification software to indicate where exons and non-coding features may be. These programs often rely on the use of gene models in certain species to offer a "blueprint" of what the gene structure of a particular organism looks like. Based on the models, gene ID tools then compare the raw sequence and return regions that resemble the exons and introns of the original examples.

There are many programs that carry out this type of analysis, and due to the test data the programs are trained on, very few return identical results. It is for this reason, that a consensus over several programs offers a better opportunity of identifying any genes and other features present.
Contents

**Eukaryotic Gene Structure** .............................. 1

- **Non-coding features** ............................. 1
  - Repetitive elements ................................ 1
  - CpG islands ....................................... 2
  - tRNA genes ........................................ 3
  - Promotor sites ..................................... 3
  - Poly-A sites ....................................... 4
  - Matrix association regions ......................... 4
  - Transcription factors ............................. 5

- **Exon/Intron identification** .................... 6

- **Codon Translation Table** ...................... 7

- **Similarity Searching** ......................... 8
  - Contamination .................................... 9

- **NIX – everything at once** .................... 9
  - Programs ........................................... 10
  - Output Display .................................. 11

**Prokaryotic Gene Identification** ............... 11

- **Artemis** .......................................... 12

**Practical** ........................................... 13
Gene Identification

NIX ............................................................. 13

Gene Identification Software ................. 13

References ......................................................... 17
Eukaryotic Gene Structure

Given a genomic sequence, it is generally a good idea to find as many features as possible and characterise the coding regions. The faster this initial stage can be carried out, the more information will be placed at the disposal of teams investigating the function of this particular region of sequence. Gene identification often involves identification of non-coding features of interest in the sequence, together with the exons and the eventual cDNA. This information can then be used to find out more about the function of this particular gene.

Often these steps are blurred into one, as when a genomic sequence is found to match the cDNA precursor to a known protein - then the exons and function may become clear at the same time. In the absence of such a match, there are software tools designed to assist during each of these stages.

Non-coding features

These will prove more or less interesting, depending on the organism of origin, and whether the expected function would represent that of a housekeeping gene, or one that is more selectively expressed.

Repetitive elements.

Repetitive elements such as Alu's are littered around the genome. It's good to know they are there, especially if the intention is to conduct experiments such as targeted gene inactivation by homologous double recombination, since too many such elements will dramatically reduce the proportion of targeted versus random integration events! Designing primers and oligo probes may also be more efficient using a masked sequence.

Repeats, such as direct tandem repeats, are also frequent in genomic DNA. Because some have putative regulatory functions, they can be found upstream of certain genes. It is always noted in the annotations when a sequence is submitted to EMBL that repeat/repetitive regions have been found.

Repetitive elements, such as the Alu family of short repeats are frequently classed as coding sequences by exon-finding programs. To stop these programs getting confused it is common practice to mask these out, replacing them by N's and there are several software options for doing this.
RepeatMasker was written at Washington University, and is available on a web server\(^1\) and as an integrated part of other software applications, such as some BLAST tools. DNA repeats and low complexity regions are masked out of sequences and replaced by Ns. Regions that may be masked out include: SINEs Alu and MIR; LINEs L1 and L2; LTRs, MaLR and Retroviral and DNA transposons. Small nuclear RNA repeats and EST sequence matches are also filtered out. Repeats from mammalian organisms are aligned against those regions available in Repbase libraries\(^2\). Repeat regions for other organisms are less well defined, and this may be reflected in the results of RepeatMasker.

The program uses optimally designed matrices to align the query DNA sequence against repeat and low complexity regions using the local alignment dynamic programming algorithm of Smith and Waterman\(^3\). Alignment scoring is adjusted to reflect complexity of a repeat. These matrices are designed to optimise the alignment based on a background of varying GC (35-53\%) content and divergence of the repeat sequences (14-25\%). The default GC level is 43\%. Alignment is based on selection of a word (generally 10 bp) which is then aligned with the repeat sequences. Should the raw sequence be used to search an EST database subsequent to masking, it would be wise to choose a longer word size with which to filter repeat sequences.

The Genetic Information Research Institute (GIRI), the home of Repbase, is also host to its own masker tool, CENSOR\(^4\). This functions in a similar way to RepeatMasker and offers principally human and rodent sequences for CENSORing, together with plant and invertebrate sequence.

XBLAST - (not to be confused with BLASTX) takes the results of Blast searching a database of repeats and masks out the regions which match entries in the database from your query sequence. Fast, but sensitive to the parameters of the BLAST search you do against the database of repeat sequences.

### CpG islands

About half of all mammalian genes have a CG-rich region around their 5' end and it is assumed that all mammalian housekeeping genes have a CpG island Non-mammalian vertebrates have some CpG islands that are associated with genes, but the association gets equivocal in the farther taxonomic groups. Non-constitutively expressed genes do not always contain a CpG island, and when they do, it does not necessarily have a 5' bias. Software applications to find these islands are derived from an algorithm developed by Larson et al\(^5\) based on a study into CpG islands carried out using 375 genes, and 58 pseudogenes.
The CpGplot program in EMBOSS functions using the sliding window technique, common for dotplots and other applications. The window slides across the sequence, calculating percentage CG content, and an observed/expected ratio of CG content, where the expected ratio is the number of cytosines and guanines present in the input sequence divided by the window length. The observed value is the multiple of consecutive cytosines/guanines combinations present in the window.

If the percentage CG content exceeds 50% and the Obs/Exp ratio remains greater than 0.6 over a default region of 200 bases, a CpG island is reported in that region. These parameters may obviously be altered by the user to reflect peculiarities in a specific organism, or tissue from which the gene may have been sequenced.

trNA genes.

These are predicted extremely well by the excellent program tRNAscan-SE which can be found at the University of Washington http://www.genetics.wustl.edu/eddy/trNAscan-SE/. This program represents a combination of three algorithms\textsuperscript{vii} \textsuperscript{viii} \textsuperscript{ix}. The program recognises potential tRNA sites based on several models created from multiple sequence alignments and individual tRNA sites taken from the database.

The user selects a kingdom from which to represent the organism of origin of the DNA query sequence. This will determine which tRNA models are chosen in the program. A first pass is made searching for everything remotely resembling the tRNA site make up. The results of this search are then handed to a more stringent program, which makes a sensitive comparison between them and existing tRNA models. Using this method, both speed and sensitivity are optimised, and the incidence of false positives reduced.

Promotor sites

Prediction of these is poor with many false positives and false negatives. Often promoter finding programs are integrated into gene identification tools, and rely again on models of various promoters to compare against the query sequence. Promoter Scan uses a model based on the density of transcription factor binding sites to suggested polymerase II promoter regions.

Another approach for promoter recognition is based on the analysis of hexamer frequencies derived from promoter regions, coding regions and non-coding
regions. In general, it is very difficult to predict the promoter region due to a high heterogeneity and complexity of these functional regions, thus the results of many promoter prediction programs should be compared. A predicted promoter site upstream of a predicted gene is good supportive evidence.

**Poly-A sites**

Polyadenylation is a very important early step of pre-mRNA processing. The most well-known signal involved in this process is AATAAA, located 15-20 nucleotides upstream from the poly-(A) site (site of cleavage). Real AATAAA signals can differ from AATAAA consensus sequence. The most frequent natural variant, ATTAAA, is nearly as active as the canonical sequence. Analysis of neighbouring bases showed that certain other bases can be important for AATAAA recognition. An additional signal with consensus YGTGTTY (diffusive GT-rich sequence) was revealed in region from 20 to 30 nucleotides downstream of poly-(A) site.

This information has been used in a variety of guises to identify potential polyadenylation sites within a query sequence. A poly A site at the end of an exon-rich region is good supportive evidence that there is a gene and that it may end at the predicted poly A region. Prediction of these is poor with many false positives and false negatives. Many exon-finding programs will also attempt to predict poly-A sites, using pattern matching algorithms and weight matrices.

A weight matrix based prediction is employed in the GRAIL gene identification system. The Hamming-Clustering (used in AATAA recognition above) network technique has been also applied for poly-(A) site prediction and is available on the WebGene server (http://www.itba.mi.cnr.it/webgene) in Italy.

**Matrix association regions**

Matrix attachment regions are associated with sites which control the expression of genes. These regions span several hundred base pairs and are not characterised by a single model.

The software tools designed to search for such regions, do so once again on the basis of biological models. In this case the models are represented by several patterns known to occur at MAR sites, which are built into the program.
The program MAR-Finder\textsuperscript{xvi}, created at the Futuresoft Cooperation, uses 20 different biological patterns associated with MAR regions, such as topoisomerase II and the origin of replication. Various permutations of each co-occurrence known to occur are used, and the density of each combination represents a model to search the raw DNA sequence. This program uses a sliding window technique to search for these sites, which are, on average 650 bp in length.

The program Marscan searches for just two regions, an 8 bp (AATAAYAA) and one 16 bp (AWWRTAANWNNC) segment known to occur together in a MAR recognition site (MRS)\textsuperscript{xvi}. These elements may overlap one another, or occur one after the other. Should several 8 and 16 bp segments be found within 200 bp of each other, then only the two closest regions are reported in the output.

**Transcription factors**

In a similar vein to other programs which predict non-coding features of the genome, prediction tools rely on biological models of these sites. The problem is not finding transcription factor binding sites, but dealing with the vast numbers of false positives. This is because transcription factor binding sites are rather small in size, often only a couple of base pairs in length, and thus random matches with portions of the DNA which do not represent these sites are difficult to avoid.

The basis for the majority of transcription factor binding site prediction programs is the TRANSFAC database\textsuperscript{xviii}, which details all transcription factor binding sites thus far identified. This is free to all academic research groups, although commercial organisations are asked to obtain a licence. There are other transcription factor databases, but this is generally considered the most complete. Tools may then just take the sequences found in the database, and match them againstquery DNA sequences to identify potential binding sites. This is the approach used in the EMBOSS program tfscan.

Other, more sophisticated tools then pass the results of this string search into a sequence alignment, which will score matches according to a weight matrix\textsuperscript{xix} defined by multiply aligning transcription factor binding sites. This is the approach used by TESS (Transcription Element Search System)\textsuperscript{xx} and is similar to the method used by MatInspector\textsuperscript{xxi}.
Exon/Intron identification

The main object of analysing a genomic sequence is to find the coding regions. In eukaryotic genomes, this entails searching for exons and their exact splice site boundaries.

There are several gene finding programs available, and all rely on an heuristic method of comparison between the query sequence and biological models based on known exonic and intronic structures together with other non-coding features, that constitute a gene. They attempt to predict genes using a model of gene structure from the promoter site through the first exon, introns, internal exons, to the last exon and the poly-A site. This structure also includes alternate exon/intron boundaries and their possible regulatory sites. Gene identification program aim to distinguish regions and features which compare favourably with the programmed pattern, and then attempts to order the segments in the most biologically probable fashion.

Most, if not all, current exon detection algorithms are notorious for missing very short exons, hence predictions tend to be less reliable for the commonly shorter 5’ and 3’ exons of intron-containing genes. These delimiting exons also tend to be separated from internal exons by longer than average introns, thus reducing accuracy even further. Some early gene model predicting programs expect only one gene in the sequence. This means they can get things wrong when asked to inspect large chunks of genomic DNA containing multiple independent genes.

Although good results have been obtained with a variety of computational approaches, the problem of gene structure prediction has not yet been completely solved and there are still many false positive and false negative predictions. The best strategy at present is probably to compare the results of many programs.

Genscan is possibly one of the most widely used gene prediction programs. Developed by Chris Burge at Stanford University (http://genes.mit.edu/GENSCAN.html), it is freely available for academic use. Commercial users need to obtain a licence. It relies very much on a probabilistic scores for essential gene structural properties, to compare query sequence features to those it has been "trained" on - namely real coding regions. It also uses Hidden Markov Models based on sound statistical models of genetic features. Genscan has proved accuracy of gene identification above a number of other tools and is favoured by the developers at EnsEMBL as one of their software analysis applications.
A further source of gene identification programs are those of Victor Solovyev who has created a multitude of programs for the Sanger Institute in the UK and Baylor College of Medicine in the USA. A full range of programs are available on the Sanger web serve (http://genomic.sanger.ac.uk/), including FSGENEH\textsuperscript{xxv}, a program using statistical probabilities based on hidden Markov models to identify genetic components. Various programs form part of the Computational Genomics Group (CGC) suite of genomic applications.

A further product, GRAIL\textsuperscript{xxvi}, uses sliding windows to locate and assess potential genetic features based on neural networking.

## Codon Translation Table

This is known as the standard codon usage table and is used primarily for all eukaryotic sequences as default, unless otherwise specified (for example in yeast).

The small $i$ next to the amino acid three letter code represents initiation sites. AUG is the primary one, with two alternative Leucine codons representing alternative initiation signals.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
<th>Codon</th>
<th>Amino Acid</th>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>F Phe</td>
<td>TCT</td>
<td>S Ser</td>
<td>TAT</td>
<td>Y Tyr</td>
</tr>
<tr>
<td>TTC</td>
<td>F Phe</td>
<td>TCC</td>
<td>S Ser</td>
<td>TAC</td>
<td>Y Tyr</td>
</tr>
<tr>
<td>TTA</td>
<td>L Leu</td>
<td>TCA</td>
<td>S Ser</td>
<td>TAA</td>
<td>* Ter</td>
</tr>
<tr>
<td>TTG</td>
<td>L Leu</td>
<td>TCG</td>
<td>S Ser</td>
<td>TAG</td>
<td>* Ter</td>
</tr>
<tr>
<td>CTT</td>
<td>L Leu</td>
<td>CCT</td>
<td>P Pro</td>
<td>CAT</td>
<td>H His</td>
</tr>
<tr>
<td>CTC</td>
<td>L Leu</td>
<td>CCC</td>
<td>P Pro</td>
<td>CAC</td>
<td>H His</td>
</tr>
<tr>
<td>CTA</td>
<td>L Leu</td>
<td>CCA</td>
<td>P Pro</td>
<td>CAA</td>
<td>Q Gin</td>
</tr>
<tr>
<td>CTG</td>
<td>L Leu</td>
<td>CGG</td>
<td>P Pro</td>
<td>CAG</td>
<td>R Arg</td>
</tr>
<tr>
<td>ATT</td>
<td>I Ile</td>
<td>ACT</td>
<td>T Thr</td>
<td>AAT</td>
<td>N Asn</td>
</tr>
<tr>
<td>ATC</td>
<td>I Ile</td>
<td>ACC</td>
<td>T Thr</td>
<td>AAC</td>
<td>N Asn</td>
</tr>
<tr>
<td>ATA</td>
<td>I Ile</td>
<td>ACA</td>
<td>T Thr</td>
<td>AAA</td>
<td>K Lys</td>
</tr>
<tr>
<td>ATG</td>
<td>M Met i</td>
<td>ACG</td>
<td>T Thr</td>
<td>AAG</td>
<td>K Lys</td>
</tr>
<tr>
<td>GTT</td>
<td>V Val</td>
<td>GCT</td>
<td>A Ala</td>
<td>GAT</td>
<td>D Asp</td>
</tr>
<tr>
<td>GTC</td>
<td>V Val</td>
<td>GCC</td>
<td>A Ala</td>
<td>GAC</td>
<td>D Asp</td>
</tr>
<tr>
<td>GTA</td>
<td>V Val</td>
<td>GCA</td>
<td>A Ala</td>
<td>GAA</td>
<td>E Glu</td>
</tr>
<tr>
<td>GTG</td>
<td>V Val</td>
<td>CGG</td>
<td>A Ala</td>
<td>GAG</td>
<td>E Glu</td>
</tr>
</tbody>
</table>

Real sequences may contain artificial nucleotide substitutions, insertions and deletions, since many sequencing strategies are error-prone. This fact should be taken into account in sequence analysis. Several algorithms were developed for the analysis of frame-shift errors\textsuperscript{xxvii} \textsuperscript{xxviii} based on statistical properties of coding sequences. The accuracy of error...
correction is not very high, but an improvement in prediction accuracy for cases with a high frame-shift rate was shown. Error identification can be significantly improved by using information on homologous proteins. Error correction techniques should be used carefully, since there are many pseudogenes in eukaryotic genomes and straightforward computer correction should be confirmed by experimental procedures.

**Similarity Searching**

Once a prediction has been made, there are several ways of verifying the prediction before it must be confirmed in the laboratory. The most obvious of these is to conduct a similarity search with DNA sequences already available in the databases. Whilst a hit may not always be good news, in particular if this gene were felt to be novel, at least it would suggest the success of the prediction process. Alternatively, partial fragments of your sequence may display local alignment with regions in other genes. A match to a database entry may offer information on your genomic region. There is a good chance of identifying the gene, exon start and end positions as well as the possible function of the gene.

Other indications may be gained from:

- **EST** - suggests transcription of the sequence. Beware, the database contains at least 1% genomic contamination and junk.
- **HTG** - recent high throughput project results - shows that other groups are working on this region!
- **GSS** - Genome survey sequences - random short segments of genome from many sequences - may indicate that someone has further sequences on this region held privately, as well as possible clone libraries from their species.
- **STS** - Pairs of markers that may be used for mapping.

A search against protein databases using a translated version of your sequence could also provide valuable information, not least if part of your translation matches a recognised protein or domain family.
Gene Identification

- Swissprot - The best (most reliable and well documented) protein database. Annotated by humans!
- TREMBL - A database of automatically translated putative proteins found in the EMBL database.

It is estimated that approximately 30% of all proteins in the Swissport database have missing exons, or introns that have been erroneously translated as part of the protein entry and 10 to 20% of proteins may be mis-identified in their annotation.

Beware of finding matches with pseudogenes or matches to collagen in the database if the query sequence contains proline-rich regions.

Contamination

It is always worth doing a BLAST search against databases of vector sequence and the bacterial or yeast genome which played host to the gene of interest. A hit here could indicate contamination at one of the earlier purification stages, and further analysis would be necessary to clarify this.

NIX – everything at once

There are individual programs to locate and identify each of these features in a raw nucleotide sequence, but as a first pass, it is useful to gain an overview of what one might expect from the sequence.

Nix is intended as a tool to aid the identification of interesting regions in genomic or transcribed nucleic acid sequences. There are many useful computer tools to do this, none of these are 100% accurate and so it is useful to be able to compare the results of many programs which use different methods and so will fail with different things. Viewing the results of many such programs side by side makes it easy to see when many programs have a consensus about a feature.

The NIX system should select reasonable defaults for the programs based on the sequence species of origin and its size. This makes a vastly simplified interface for running a dozen or so programs on the specified sequence. The disadvantage of this, however, is that the user cannot tailor any of the parameters to a particular sequence, but should the NIX results display any interesting results, programs may be run individually on the sequence.
The submitted sequence is masked for repeat regions using Washington University's repeatmasker program and BLAST searches are started using the masked sequence against the databases: Ecoli, EST, EMBL (minus the sts, est, gss sections) using TBLASTX, TrEMBL, SwissProt. The Expect value cutoff is set to 0.1. Up to 1000000 alignments can be returned.

For transcribed sequences, GRAIL is used for exon identification. Genomic sequences are handled as above, and they are also subject to the following gene identification programs:

- **GeneFinder** uses log likelihood ratios to score potential gene candidates.
- **GeneMark** uses statistical models based on the back translation of proteins to assess how likely a region would be to code for a protein.
- **GenScan** uses HMM techniques to assess the probability of specific regions representing genetic features.
- **Fex** finds potential exons using hidden Markov model prediction.
- **Hexon** finds exons based on the linear discriminant function of donor (GT) and acceptor (AG) splice sites.
- **Fgene** Once again, (linear) discriminant functions are used to assess potential exons. Dynamic programming is then used to place these in a suitable order.
- **MZEP** This program uses quaternary discriminant analysis of nine functions to predict exons.

tRNA sites are also predicted by trnascan-se.
Output Display

The NIX output is displayed as a graphical view. Features identified on the leading DNA strand of the query sequence are shown on the top half of the display, the bottom half represents the results of the anti-sense strand. The view is halved by a representation of the query sequence as a thick green line.

Programs are colour codes according to similarity, and the results appear as triangular blobs, and rectangular boxes in the same colours. A brighter hue indicates a good or excellent hit, whilst a darker definition indicates a marginal, or poor identification. For browsers with a java facility enabled, placing the mouse over the resulting display will identify the result in a small box above the graphical output. Selection of a result with the left mouse button (single click) will open up another window in which the raw output of the program results are displayed.

Each program used in NIX is documented and this can be accessed by selecting the program name on the left hand side of the results display. Documentation is opened in the same window as the display.

**Prokaryotic Gene Identification**

It is a biological fact that prokaryotes differ in many ways to eukaryotes. Although much bioinformatics revolves around the identification of features in a eukaryotic genome, effort is also directed towards bacteria, not least the parasites that infect our animals and crops.

Prokaryote genes, unlike their eukaryotic counterparts, are, co-linear. Coding sequences are (generally) free from introns, and the open reading frame codes directly for a protein with no internal splicing. Many bacterial genes are found as components of operons, controlled at a single promoter site.

Promoter sites are characterised by regions at -35 and -10 (Pribnow Box) upstream of the initiation site. Multiple alignment of these sequences has defined a consensus series of TTGACA and TATAAT nucleotides respectively. This offers a basis on which prokaryotic gene identification programs can build. Operator regions, adjacent to promoter sequences are not always conserved, and cannot as easily be used as a search model.

Protein translation is initiated by the attachment of the ribosome to the mRNA at a ribosome binding sequence (RBS). This is characterised by the purine rich Shine Dalgarno sequence (AGGAGG) several bases upstream from the AUG initiation site.
Bacterial transcription is often terminated by a stem loop structure in the mRNA. This feature is represented by a region of dyad symmetry in the DNA sequence, which again, may be recognised by computational tools.

Thus it is obvious that programs which are used to identify genetic features in prokaryotes must adhere to a different set of rules to those used in tools for eukaryotic gene identification.

The majority of software tools available to search for eukaryotic genes can also be used for the annotation of bacterial genes. This must be specified by the user before running the program.

**Artemis**

One of the free tools for viewing and annotation of DNA sequence is Artemis^{cov}. It has been programmed to function particularly well for the identification of prokaryotic feature and is the main tool used in the analysis of microbial genomes at the Sanger Institute\(^1\).

It will display features of a DNA sequence in all six frames, and calculate GC content and other physical characteristics on either the whole sequence, or CDS region. Results of further analyses are displayed on top of these sequences.

AS with Jemboss, Artemis is written in Java, which must be available on the computer for the program to function.

\(^1\) http://www.sanger.ac.uk
Gene Identification

Practical

NIX

NIX can take up to a day to run if it is to be done accurately, but, depending on the queue and the sequence length, most programs will perform fairly rapidly.

Go to the HGMP homepage at http://www.hgmp.mrc.ac.uk and select "NIX" from the bottom of the page. Follow the "Run NIX now!" link and paste the genomic sequence into the relevant field. Give your sequence a description and type in your email address.

Have a look at the NIX images of the sequence, which show features found by the various programs. The features are represented by graphical blocks, which link to the raw output of results.

Use the help pages (available by clicking on the name of the program on the left hand side of the display) to understand what the various programs have tried to predict. These help pages can also assist you in making the most of the raw outputs.

Take a look at the BLAST outputs of the sequence. How well does this compare with what you saw after your earlier BLAST search?

Does a consensus appear for the predicted exons in pax6 gene? What about a clear gene structure (CpG island, promotor, exons, polyA)?

Gene Identification Software

A relatively recent advance in bioinformatics has been the development of software specialised in gene identification (gene ID). These programs are used to recognise and extract the functional genetic information encoded in novel DNA sequences. Rather than collecting data for use in the laboratory, gene ID software assists investigators in the characterisation of such diverse genetic

---

2 If you wish to analyse your own sequence, select "Run NIX now!" and paste it into the entry field.

---

Human Genome Mapping Project Resource Centre  http://www.hgmp.mrc.ac.uk
features as promoters, splice sites, coding versus non-coding regions, polyadenylation signals etc.

The individual features of a DNA sequence can be integrated to form a complete gene model. The efficiencies of the various gene ID algorithms are closely tied to our current understanding of molecular genetics. Although many different approaches are used to pick out specific features in unknown DNA sequences, all rely on past experience of the programmer (or program) with the feature under study. For example, when looking for splice sites in genomic DNA, some programs use the best currently accepted consensus (e.g. AG dinucleotide at intron/exon boundary) to scan the unknown sequence, while other programs are trained with a set of splice sites in known sequences before they are asked to evaluate an unknown sequence (neural network approach). Indeed, as with typical biology experiments, gene ID applications often have non negligible false positive and false negative rates in their predictions. The specific predictions (true positive) rates fluctuate between 50% and 95%, depending on the programs used and the DNA sequences submitted. Recent significant advances in the development of bioinformatics tools for gene identification have taken advantage of so called Hidden Markov Models (HMM) which are based on sound statistical models (e.g. GENSCAN).

There are many GeneID programs on the WWW. After using applications like NIX to gain an overview of what is happening, you may then want to look in more detail and start tweaking parameters. One of the best single programs for \textit{de novo} gene prediction is GENSCAN, which uses HMM techniques (see next section).

Go to the HGMP homepage and follow the link to the "The Genome Web" at the bottom of the page. Then follow the subsequent "Nucleic Acids"; "Gene Identification" and "Genscan". Scroll down the page until you reach the entry fields for your input sequence. Accept the default for organism and cutoff, and enter a description for your sequence. Select "Predicted CDS and peptides". Copy and paste your pax6\_genomic.fasta sequence from the WWW\(^3\) or Jemboss file manager into the data entry field. Enter your email address and press the "Run GENSCAN" button.

Take a look at the results. The predicted exons are displayed in a table where "intr" signifies an internal exon, "Term" signifies a terminal exon, and "PlyA"

\(^3\) This is available by following the "WWW menu" link from the HGMP homepage. Select "Utilities", File Management" to the "Simple WWW file manager". Click on your file selection, and it will be displayed to you in the web browser from where you can copy it.
describes the position of a poly adenosine tail. "Sngl" defines a single exon gene, and "Prom" a promoter region.

GENSCAN is one of the gene finding programs that will recognise if there is more than one potential gene in a sequence, and will also make a good attempt at recognising partial genes. Each predicted gene is numbered (in this example, there are three predicted genes) and each of the features of that gene is also numbered (for example, 1.01, 1.02, and so on). The gene number is displayed, to the left of the feature definition (Type) followed by the strand (S) on which the feature has been found. The start (Begin) and end (End) positions of the features are then displayed, together with the length (Len).

Various scores are presented, and you can read about exactly what they are on the GENSCAN web page. Each predicted exon is assigned a probability, P, which is "the estimated probability under GENSCAN’s model of genomic sequence structure that the exon is correct. This probability depends in general on global as well as local sequence properties." Trials with sets of test data suggest that very high probability exons (P > 0.99) are nearly always correct, those with 0.50 < P < 0.99 are correct most of the time, while those with P < 0.50 are not reliable. GENSCAN has been designed primarily for vertebrate sequences and may be less accurate for non-vertebrates.

Predicted genes/exons:

<table>
<thead>
<tr>
<th>Gn.Ex</th>
<th>Type</th>
<th>S</th>
<th>Begin</th>
<th>...End</th>
<th>Len</th>
<th>Fr</th>
<th>Ph</th>
<th>I/Ac</th>
<th>Do/T</th>
<th>CodRg</th>
<th>P.....</th>
<th>Tscr..</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.01</td>
<td>Intr</td>
<td>+</td>
<td>1078</td>
<td>1194</td>
<td>117</td>
<td>2</td>
<td>0</td>
<td>71</td>
<td>52</td>
<td>101</td>
<td>0.610</td>
<td>4.32</td>
</tr>
<tr>
<td>1.02</td>
<td>Intr</td>
<td>+</td>
<td>1380</td>
<td>1510</td>
<td>131</td>
<td>1</td>
<td>2</td>
<td>141</td>
<td>58</td>
<td>68</td>
<td>0.979</td>
<td>8.59</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.99</td>
<td>Intr</td>
<td>+</td>
<td>5421</td>
<td>5610</td>
<td>528</td>
<td>2</td>
<td>0</td>
<td>371</td>
<td>36</td>
<td>280</td>
<td>0.586</td>
<td>-5.16</td>
</tr>
</tbody>
</table>

---

Human Genome Mapping Project Resource Centre
http://www.hgmp.mrc.ac.uk
Gene Identification

Three genes have been predicted (defined by numbers 1, 2 and 3 on the left hand side of the results⁴). Does this correspond to what you have seen in the NIX analysis?

Number of Exons (gene 1)  

Start and finish of each

Does this correspond with the information in Ensembl? Why not⁵? What about the information from your NIX results?

---

⁴ The numbering is reversed for predictions 2 and 3 as they have been found on the anti-sense strand.

⁵ Ensembl uses Genscan as only one of it’s prediction tools. Genes are built using several sources of information. Also, some of the exons have been found on another clone, the sequence of which is no contained in Z83307.
References


http://www.girinst.org/ Registration is necessary, but the libraries are free to non-profit making organisations

Identification of common molecular subsequences

CENSOR - a program for identification and elimination of repetitive elements from DNA sequences.
http://www.girinst.org/Censor_Server-Data_Entry_Forms.html

CpG islands as gene markers in the human genome

tRNA scan-SE: A program for improved detection of transfer RNA genes in genomic sequence

RNA sequence analysis using covariance models

Identifying potential tRNA genes in genomic DNA sequences.

Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions.

Predicting Pol II promoter sequences using transcription factor binding sites.
http://www.molbiol.ox.ac.uk/promoterscan.htm
The prediction of vertebrate promoter regions using differential hexamer frequency analysis.

xii Proudfoot, N., (1991) *Cell*. 64 (4) 671-4
Poly(A) signals.

Hamming-Clustering method for signals prediction in 5’ and 3’ regions of eukaryotic genes

The consensus sequence YGTGTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3’ termini.

Discovering and understanding genes in human DNA sequence using GRAIL.

xvi http://www.futuresoft.org/index.html

A bipartite sequence element associated with matrix/scaffold attachment regions.

TRANSFAC: a database on transcription factors and their DNA binding sites.
http://transfac.gbf.de/TRANSFAC/

Modelling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding.

xx Schug, J., Overton, G.C., (1997) "TESS: Transcription Element Search Software on the WWW" in Technical Report CBIL-TR-1997-1001-v0.0, of the Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania
http://www.cbil.upenn.edu/tess

Matind and Matinspector - New Fast and Versatile tools for Detection of Consensus Matches in Nucleotide-Sequence Data
Prediction of complete gene structures in human genomic DNA

Evaluation of gene structure prediction programs

The Ensembl genome database project  
http://www.ensembl.org

Salamov A.A., Solovyev V.V. (1999), unpublished data.  
CGG WEB server: http://genomic.sanger.ac.uk/gf/gf.shtml

Xu, Y., Manesh B. Shah, J. Einstein, R., Parang, M., Snoddy, J., Petrov, S., Olman, V., Zhang, G., Mural, R.J., Uberbacher, E.C.  
GRAIL and GenQuest Sequence Annotation Tools, *Third Pacific Symposium on Biocomputing*, Santa Fe, NM, November 1997  
http://grail.lsdornl.gov/grailexp/

Correcting sequencing errors in DNA coding regions using a dynamic programming approach.

A frameshift error detection algorithm for DNA sequencing projects.

Wilson, C., Hilyer, L., and Green, P., unpublished,  
http://repeatmasker.genome.washington.edu/cgi-bin/Genefinder

GeneMark: Parallel Gene Recognition for both DNA Strands  
http://opal.biology.gatech.edu/GeneMark/

Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames

