© 1992 Oxford University Press · Nucleic Acids Research, Vol. 20, No. 13 3453–3462

# The prediction of exons through an analysis of spliceable open reading frames

Gordon B.Hutchinson* and Michael R.Hayden
Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Received February 24, 1992; Revised and Accepted May 21, 1992

## ABSTRACT

**We have developed a computer program which predicts internal exons from naive genomic sequence data and which will run on any IBM-compatible 80286 (or higher) computer. The algorithm searches a sequence for 'spliceable open reading frames' (SORFs), which are open reading frames bracketed by suitable splice-recognition sequences, and then analyzes the region for codon usage. Potential exons are stratified according to the reliability of their prediction, from confidence levels 1 to 5. The program is designed to predict internal exons of length greater than 60 nucleotides. In an analysis of 116 genes of a training set, 384 out of 441 such exons (87.1%) are identified, with 280 (63.5%) of predictions matching the true exon exactly (at both 5′ and 3′ splice junctions and in the correct reading frame), and with 104 (23.6%) exons matching partially. In a similar analysis of 14 genes in a test set unrelated to the genes used to generate the parameters of the program, 70 out of 80 internal exons greater than 60 bp in length are identified (87.5%), with 47 completely and 23 partially matched. SORFs that partially match true internal exons share at least one splice junction with the exon, or share both splice junctions but are interpreted in an incorrect reading frame. Specificity (the percentage of SORFs that correspond to true exons) varies from 91% at confidence level 1 to 16% at confidence level 5, with an overall specificity of 35–40%. The output displays nucleotide position, confidence level, reading frame phase at the 5′ and 3′ ends, acceptor and donor sequences and scoring statistics and also gives an amino acid translation of the potential exon. SORFIND compares favourably with other programs currently used to predict protein-coding regions.**

## INTRODUCTION

Improved sequencing technologies and the initiatives of the Human Genome Project are generating large amounts of naive DNA sequence, which has motivated the search for efficient computer algorithms to identify coding regions in genomic DNA. When starting with a contiguous cloned DNA segment, the identification of genes usually proceeds by looking for expressed sequences in cDNA libraries. However, differing tissue specificity and differences in the temporal expression of genes can lead to their under-representation in these libraries. Moreover, it is often difficult to obtain full length cDNA transcripts. In the early stages of the characterization of a cloned genomic DNA fragment, when only a partial sequence of a gene may be present, computer algorithms that identify potential coding regions can serve to focus efforts on those DNA segments that are more likely to represent exonic sequences. In addition, translations of these sequences into their amino acid equivalents can significantly enhance database searches for homologous proteins.

Computer methodologies for identifying coding regions can be classified into two types (1). The first, gene search by signal, relies upon the identification of short sequences such as those characteristic of splice junctions or promoters. Matrix methods for scoring these consensus sequences are commonly used (2, 3, 4), and approaches utilizing neural nets are becoming more widespread( 5, 6). A second methodology, which can be termed gene search by content, looks at long segments of DNA to see if they resemble coding sequence. Examples of this include codon usage and preference measurements (1), k-tuple frequency analysis (7), local compositional complexities (8) and neural net approaches (e.g. the Gene Recognition and Analysis Internet Link, GRAIL project at Oakridge National Laboratories (9). There have now been several attempts to combine the two methodologies. Gm (*Gene Modeler*) (10), uses a log-likelihood method to score splice junctions and measures AT versus GC richness (among other parameters) in and around open reading frames to predict gene assemblies. Discrimination energy (3, 4) in conjunction with a codon usage algorithm (11) has been used to predict mammalian exon assemblies (12), but without a statistical analysis of sensitivity and specificity. Finally, in the program GeneId (13), a weighted profile of initiation codons, acceptor and donor sites to initially select possible exons is used. The result is filtered by rejecting exons based upon 24 variables of nucleotide fraction and codon position correlations.

In this paper, we present an alternative method which combines the discrimination energy as described by Berg and von Hippel with three measures of codon usage and predicts internal exons at 5 confidence levels. Our program, named SORFIND, presents

* To whom correspondence should be addressed at: Canadian Genetic Diseases Network, 2125 East Mall, Vancouver, BC V6T 1Z4, Canada

the user with exon predictions aimed at identifying human genes when only partial sequence is available. The program is able to identify 87% of internal exons, with a specificity that varies from 91.5% for confidence level 1 predictions to 15.8% for confidence level 5 predictions, for an overall specificity of 39%. The predicted exons can be used to design more specific probes of cDNA libraries, and its amino acid sequence predictions allow more specific database searches for protein sequence homology. The program can analyze a 28 kb sequence in less than 4 minutes on an 80286 IBM-compatible microcomputer with co-processor installed.

## MATERIALS AND METHODS

### Creation of data sets

Sequence files were initially extracted from GenBank 67 which met the following criteria: i) the locus name starts with 'HUM', ii) the definition line contains the word 'complete', and iii) at least 3 feature lines contain the words 'exon' or 'pept', or at least 2 lines contain the words 'IVS' or 'intron'. This created a subset of 190 loci likely to contain only human genes with complete coding sequences and at least one internal exon. It was then necessary to further reduce that data set in order to eliminate genes which might have unusual or aberrant splicing and to minimize any bias that might be introduced due to the over-representation of certain gene families in GenBank. Each annotation was examined in detail, and loci were removed which contained multiple genes (5), alternate splicing (4), no introns (2), duplicate genes (14), pseudogenes (8), mutant alleles (4) and segmented entries (3). A further 18 genes were identified by an early version of the program as being similar enough to contain exons with identical acceptor and donor splice junctions. This group, including several major histocompatibility genes, was also removed. The resulting 132 genes in the data set were used to create the codon usage table discussed below. Subsequent to this, a further 16 genes were found to be unsuitable, either because they did not contain internal exons or because their size ( >30 kb) resulted in memory allocation problems for the program. This final training data set, containing 116 entries, is listed in Table 1a, along with sequence length, number of exons and exon density, defined as the proportion of nucleotides that the feature table classifies as exonic. Following development of the program, a second, unbiased data set was required to evaluate the program's performance. A similar procedure was followed, this time using the GenBank 69 release. A further 14 genes were thus identified, and are listed in Table 1b.

### Codon usage table

Codon usage is one of several statistical properties of protein coding regions that can be used to identify probable exons in an unknown sequence (1). It can be defined as the frequency that a given trinucleotide appears in frame within the coding sequence of a particular protein or collection of proteins. It is thus dependent upon both the amino acid composition, which can be biased by the type of protein chosen for study, and upon the codon preference of the organism. The 132 loci of the early training set were analyzed by a program which translated the coding region of each entry, based upon the CDS feature line in the GenBank file. The program ensured that the only stop codon present was the last codon. A customized codon usage table was thus created on the basis of these 41595 codons. The frequencies were multiplied by 61 (the number of amino acid codons), minimizing the number of decimal places required in subsequent displays. The natural logarithm was taken (so that values could be added rather than multiplied). Stop codons were assigned an average value, so that comparisons with adjacent non-coding regions would not be biased. Table 2 displays the result.

### Splice junction scoring method

We have adopted the scoring methodology described by Penotti (14). He sampled 764 pairs of human pre-mRNA exon-intron and intron−exon boundaries and scored them based upon the discrimination energy defined by Berg and von Hippel (3, 4). By this calculation, a sequence which is identical to the consensus receives a score of zero, while increasing departure from the consensus is indicated by a positive increasing value. The upper bound of this score (the worst possible match) is 30.1 for donors and 42.5 for acceptors. The system is equivalent to a log-likelihood method, but avoids negative numbers and gives a fixed reference point. It also gives a meaningful score in the case of non-consensus sites (e.g., donor sites that utilize GC rather than GT).

### Description of the user interface and algorithm

The program first reads the input file and determines the sequence format. If it is a GenBank formatted file, the feature table is analyzed to determine the coding and exonic regions of the sequence and this is later used to compare results of the program with expected values. The input file may also consist of a list of loci to examine. In this 'batch mode', the program sequentially analyzes each file in the manner described above.

After reading in the sequence (ignoring digits and punctuation), the program scans it from the 5′ to 3′ direction, stopping at each AG dinucleotide. It rejects those sites with another AG less than 11 base pairs upstream. In the test set, this eliminated 10 out of 474 true acceptor sites (2.1%), but also eliminated over 4000 false sites. If accepted, the site is then scored according to the method described above, and is rejected if its score is above the threshold for acceptor sites (explained below). The position of the first downstream stop codon in each of the 3 reading frames is then noted, and all GT dinucleotides that are at least 60 bp downstream and within this window are analyzed. The score of each potential donor site is calculated, and the site is rejected if its score falls above a donor threshold value. If a given sequence segment survives the selection procedure to this point, it must consist of at least one open reading frame containing 60 nucleotides or more, bracketed by admissible splice acceptor and donor sites. Each such reading frame, with its splice junctions, we define as a spliceable open reading frame or SORF. The program then calculates three separate variables, based upon codon usage, for each SORF. The algorithm first looks upstream and sums the individual codon usage scores within a set window for the immediately adjacent 'intronic' region. It then subtracts this value from an equal window just downstream of the acceptor junction within the SORF, giving the 5′ Codon Usage Difference. The value of the window parameter can be varied. For the data given below, the window was set at either 30 codons or one third the length of the SORF, whichever was shorter. It is expected that true acceptor splice junctions will separate good and poor regions of codon usage, corresponding to exonic and intronic sequence, and a large, positive codon usage difference will result. A similar value, the 3′ Codon Usage Difference is calculated for the donor site, and a Codon Usage Average is then calculated

**Table 1.** The 116 loci of the training data set (1a) and testing data set (1b), with length of each gene in base pairs (bp), number of exons and exon density. Exon density is defined as the proportion of nucleotides in the GenBank entry that are identified in the feature table as exons.

**a Training set (taken from GenBank 67)**

| Locus | BP | Exons | Exon Density | Locus | BP | Exons | Exon Density | Locus | BP | Exons | Exon Density |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HUMA1ATP | 12222 | 4 | 0.10 | HUMFCREB | 5131 | 5 | 0.05 | HUMMT2A | 1922 | 3 | 0.10 |
| HUMA1GLY2 | 4944 | 6 | 0.12 | HUMFOS | 6210 | 4 | 0.18 | HUMOPS | 6953 | 5 | 0.15 |
| HUMACCYBA | 3657 | 5 | 0.31 | HUMGOS19B | 4788 | 3 | 0.06 | HUMOTNPI | 1338 | 3 | 0.28 |
| HUMACTGA | 3583 | 5 | 0.31 | HUMGAPDHG | 5378 | 8 | 0.19 | HUMP45C17 | 8549 | 8 | 0.18 |
| HUMAFP | 22166 | 14 | 0.08 | HUMGCB1 | 7604 | 11 | 0.21 | HUMPALD | 7616 | 4 | 0.06 |
| HUMAK1 | 12229 | 6 | 0.05 | HUMGFP40H | 4379 | 5 | 0.10 | HUMPCNA | 6340 | 6 | 0.12 |
| HUMALPI | 5291 | 11 | 0.30 | HUMGG | 6455 | 4 | 0.08 | HUMPDHBET | 8872 | 10 | 0.12 |
| HUMANFA | 2710 | 3 | 0.17 | HUMGHN | 2657 | 5 | 0.25 | HUMPGAMMG | 3771 | 3 | 0.20 |
| HUMANT1 | 5768 | 4 | 0.16 | HUMGRP78 | 5470 | 8 | 0.36 | HUMPIM1A | 6113 | 6 | 0.15 |
| HUMANT2X | 4982 | 4 | 0.25 | HUMHBQ1A | 1114 | 3 | 0.39 | HUMPNMTA | 4174 | 3 | 0.20 |
| HUMAPOA2I | 2928 | 3 | 0.10 | HUMHLL4G | 4428 | 4 | 0.11 | HUMPP14B | 8076 | 6 | 0.07 |
| HUMAPOA4A | 3613 | 3 | 0.33 | HUMHMG14A | 8882 | 6 | 0.03 | HUMPPPA | 2775 | 3 | 0.10 |
| HUMAPOAIT | 2385 | 3 | 0.34 | HUMHSP90B | 8210 | 11 | 0.26 | HUMPRCA | 11725 | 8 | 0.12 |
| HUMAPOCIA | 5375 | 3 | 0.05 | HUMHST | 6616 | 3 | 0.09 | HUMPRPH1 | 4946 | 3 | 0.10 |
| HUMAPOCII | 4340 | 3 | 0.07 | HUMI309 | 3709 | 3 | 0.15 | HUMPSAA | 7130 | 5 | 0.11 |
| HUMAPOE4 | 5515 | 3 | 0.17 | HUMIBP3 | 10884 | 4 | 0.08 | HUMPSAP | 4778 | 4 | 0.16 |
| HUMAPRTA | 2956 | 5 | 0.18 | HUMIFNINI | 5209 | 4 | 0.06 | HUMRASH | 6453 | 4 | 0.09 |
| HUMATP1A2 | 26668 | 23 | 0.11 | HUMIGFBP1 | 6480 | 4 | 0.23 | HUMREGB | 4251 | 5 | 0.12 |
| HUMATPGG | 15115 | 22 | 0.21 | HUMIL1B | 7824 | 6 | 0.10 | HUMRPS14 | 5985 | 4 | 0.08 |
| HUMATPSYB | 10186 | 10 | 0.16 | HUMIL2A | 6684 | 4 | 0.07 | HUMRPS17A | 4029 | 5 | 0.10 |
| HUMBHSD | 9404 | 4 | 0.18 | HUMIL5A | 3241 | 4 | 0.13 | HUMSAA1A | 6943 | 4 | 0.06 |
| HUMBMYH7 | 28438 | 40 | 0.21 | HUMINCP | 3716 | 3 | 0.11 | HUMSAACT | 3778 | 6 | 0.30 |
| HUMBNPA | 1922 | 3 | 0.21 | HUMIRBPG | 9711 | 4 | 0.39 | HUMSHBGA | 6087 | 8 | 0.20 |
| HUMCAD | 4306 | 12 | 0.34 | HUMKAL2 | 6139 | 5 | 0.13 | HUMSODB | 8841 | 10 | 0.16 |
| HUMCAPG | 3734 | 5 | 0.21 | HUMKER18 | 6520 | 7 | 0.20 | HUMSPRO | 5296 | 8 | 0.27 |
| HUMCKMT | 6896 | 9 | 0.18 | HUMKEREP | 5339 | 8 | 0.27 | HUMTFPB | 13865 | 6 | 0.06 |
| HUMCS1 | 2301 | 5 | 0.28 | HUMLACTA | 3310 | 4 | 0.13 | HUMTHB | 20801 | 14 | 0.09 |
| HUMCSFGMA | 3194 | 4 | 0.14 | HUMLYL1B | 4569 | 3 | 0.18 | HUMTHY1A | 2806 | 3 | 0.17 |
| HUMCTLA1A | 4751 | 5 | 0.16 | HUMMCHEMP | 2776 | 3 | 0.11 | HUMTKRA | 13500 | 7 | 0.05 |
| HUMCYC1A | 4622 | 7 | 0.21 | HUMMETIA | 2941 | 3 | 0.06 | HUMTNFA | 3633 | 4 | 0.19 |
| HUMCYP2D6 | 9432 | 9 | 0.16 | HUMMETIF1 | 2076 | 3 | 0.09 | HUMTROC | 4567 | 6 | 0.15 |
| HUMCYPIIE | 14776 | 9 | 0.10 | HUMMGPA | 7734 | 4 | 0.05 | HUMTRPY1B | 2609 | 5 | 0.32 |
| HUMDES | 8878 | 9 | 0.16 | HUMMH6 | 4361 | 6 | 0.23 | HUMTS1 | 18596 | 7 | 0.05 |
| HUMDKERB | 8815 | 8 | 0.16 | HUMMHCD8A | 7319 | 6 | 0.10 | HUMTUBAG | 4087 | 4 | 0.33 |
| HUMEDHB17 | 4845 | 6 | 0.20 | HUMMHCP42 | 5141 | 10 | 0.29 | HUMTUBBM | 3284 | 4 | 0.41 |
| HUMEF1A | 4695 | 7 | 0.30 | HUMMHDOB | 5447 | 6 | 0.15 | HUMUBILP | 3583 | 4 | 0.13 |
| HUMEMBPA | 3608 | 5 | 0.19 | HUMMHDRHA | 5724 | 4 | 0.13 | HUMVPNP | 2500 | 3 | 0.20 |
| HUMERPA | 3602 | 5 | 0.16 | HUMMHEA | 4938 | 7 | 0.22 | | | | |
| HUMFABP | 5204 | 4 | 0.08 | HUMMIS | 3100 | 5 | 0.54 | Total: | 742842 | 697 | |

**b Test set (14 additional genes from GenBank 69)**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HUMAGAL | 13662 | 9 | 0.26 | HUMFIBRA | 5943 | 5 | 0.33 | HUMPCI | 15571 | 5 | 0.14 |
| HUMALIFA | 7614 | 3 | 0.49 | HUMG6PDGEN | 20114 | 13 | 0.13 | HUMPEM | 4243 | 7 | 0.43 |
| HUMCBRG | 3326 | 3 | 0.28 | HUMHKATPC | 17201 | 22 | 0.21 | HUMSPERSYN | 7623 | 8 | 0.22 |
| HUMCHYMASE | 4019 | 3 | 0.13 | HUMIGFBP1A | 6128 | 4 | 0.25 | HUMVCAM1A | 5607 | 9 | 0.55 |
| HUMCSPA | 4791 | 5 | 0.18 | HUMNUCLEO | 10942 | 14 | 0.23 | Total: | 126784 | 110 | |

over the entire SORF. There are separate thresholds set for these three values, and if a SORF falls below threshold for two or more, it is rejected.

SORFs surviving the filtration procedure to this stage may be in conflict with others, either by overlapping them, or by being less than a minimum distance away (corresponding to the expected minimum length of an intron). Results presented here used a minimum intron length of 70, as few human introns are smaller than this. A mediation procedure identifies those SORFs that are in conflict, and passes them on to an arbitration step, which chooses the best candidate among the competing SORFs based upon a score which linearly combines the five values previously mentioned.

Following this, the sequence is partitioned, so that further analysis will avoid regions already containing a successful SORF and its surrounding minimum introns. Depending upon the number of confidence levels requested by the user, the threshold levels are then incrementally relaxed, and the procedures above repeated. For example, upon completion of the confidence level 1 scan (which was completed with threshold filtration values set to the mean of ea^h variable's distribution), the thresholds are relaxed by adding or subtracting a fraction of the standard

**Table 2.** Human Codon Usage Statistics. Statistics were compiled from 132 loci meeting the initial criteria of the data set (see text). The final 116 loci in the training set are a subset of this group. Amino acid, codon, and the number of occurrences of that codon as well as the usage frequency are listed. This frequency was then multiplied by 61 (the number of codons representing amino acids) and the natural logarithm was taken. Negative numbers therefore reflect codons used less frequently than expected by random chance.

| Amino Acid | Codon | Number | Usage Freq. | ln(61xFreq) | Amino Acid | Codon | Number | Usage Freq. | ln(61xFreq) |
|---|---|---|---|---|---|---|---|---|---|
| Stop | tga | 65 | n/a | -0.218 | Met | atg | 974 | 0.023 | 0.360 |
| Stop | taa | 34 | n/a | -0.218 | Asn | aac | 902 | 0.022 | 0.283 |
| Stop | tag | 33 | n/a | -0.218 | Asn | aat | 500 | 0.012 | -0.307 |
| Ala | gcc | 1614 | 0.039 | 0.865 | Pro | ccc | 906 | 0.022 | 0.287 |
| Ala | gct | 781 | 0.019 | 0.139 | Pro | cct | 563 | 0.014 | -0.188 |
| Ala | gca | 502 | 0.012 | -0.303 | Pro | cca | 480 | 0.012 | -0.348 |
| Ala | gcg | 396 | 0.010 | -0.540 | Pro | ccg | 289 | 0.007 | -0.855 |
| Cys | tgc | 659 | 0.016 | -0.031 | Gln | cag | 1525 | 0.037 | 0.808 |
| Cys | tgt | 311 | 0.008 | -0.782 | Gln | caa | 325 | 0.008 | -0.738 |
| Asp | gac | 1356 | 0.033 | 0.691 | Arg | cgc | 688 | 0.017 | 0.012 |
| Asp | gat | 751 | 0.018 | 0.100 | Arg | cgg | 510 | 0.012 | -0.287 |
| Glu | gag | 2218 | 0.053 | 1.183 | Arg | agg | 455 | 0.011 | -0.401 |
| Glu | gaa | 866 | 0.021 | 0.242 | Arg | aga | 294 | 0.007 | -0.838 |
| Phe | ttc | 1057 | 0.025 | 0.442 | Arg | cga | 228 | 0.005 | -1.092 |
| Phe | ttt | 496 | 0.012 | -0.315 | Arg | cgt | 197 | 0.005 | -1.238 |
| Gly | ggc | 1265 | 0.031 | 0.621 | Ser | agc | 828 | 0.020 | 0.197 |
| Gly | ggg | 714 | 0.017 | 0.049 | Ser | tcc | 794 | 0.019 | 0.155 |
| Gly | gga | 509 | 0.012 | -0.289 | Ser | tct | 415 | 0.010 | -0.493 |
| Gly | ggt | 418 | 0.010 | -0.486 | Ser | tca | 264 | 0.006 | -0.946 |
| His | cac | 613 | 0.015 | -0.103 | Ser | agt | 250 | 0.006 | -1.000 |
| His | cat | 312 | 0.008 | -0.779 | Ser | tcg | 231 | 0.006 | -1.079 |
| Ile | atc | 1164 | 0.028 | 0.538 | Thr | acc | 1069 | 0.026 | 0.453 |
| Ile | att | 513 | 0.012 | -0.281 | Thr | aca | 442 | 0.011 | -0.430 |
| Ile | ata | 157 | 0.004 | -1.465 | Thr | act | 399 | 0.010 | -0.533 |
| Lys | aag | 1680 | 0.041 | 0.905 | Thr | acg | 312 | 0.008 | -0.779 |
| Lys | aaa | 660 | 0.016 | -0.029 | Val | gtg | 1429 | 0.034 | 0.743 |
| Leu | ctg | 2306 | 0.056 | 1.222 | Val | gtc | 705 | 0.017 | 0.037 |
| Leu | ctc | 923 | 0.022 | 0.306 | Val | gtt | 320 | 0.008 | -0.753 |
| Leu | ttg | 394 | 0.010 | -0.545 | Val | gta | 218 | 0.005 | -1.137 |
| Leu | ctt | 322 | 0.008 | -0.747 | Trp | tgg | 526 | 0.013 | -0.256 |
| Leu | cta | 200 | 0.005 | -1.223 | Tyr | tac | 784 | 0.019 | 0.143 |
| Leu | tta | 103 | 0.002 | -1.887 | Tyr | tat | 381 | 0.009 | -0.579 |

deviation. The confidence level 2 scan is then conducted with these new threshold values, but without re-scanning those regions with a previously identified SORF. As the algorithm progresses through subsequent confidence levels, the thresholds of each variable are relaxed further by the same amount, so that by confidence level 5, practically all true exons should have been identified. The algorithm is represented schematically in figure 1.

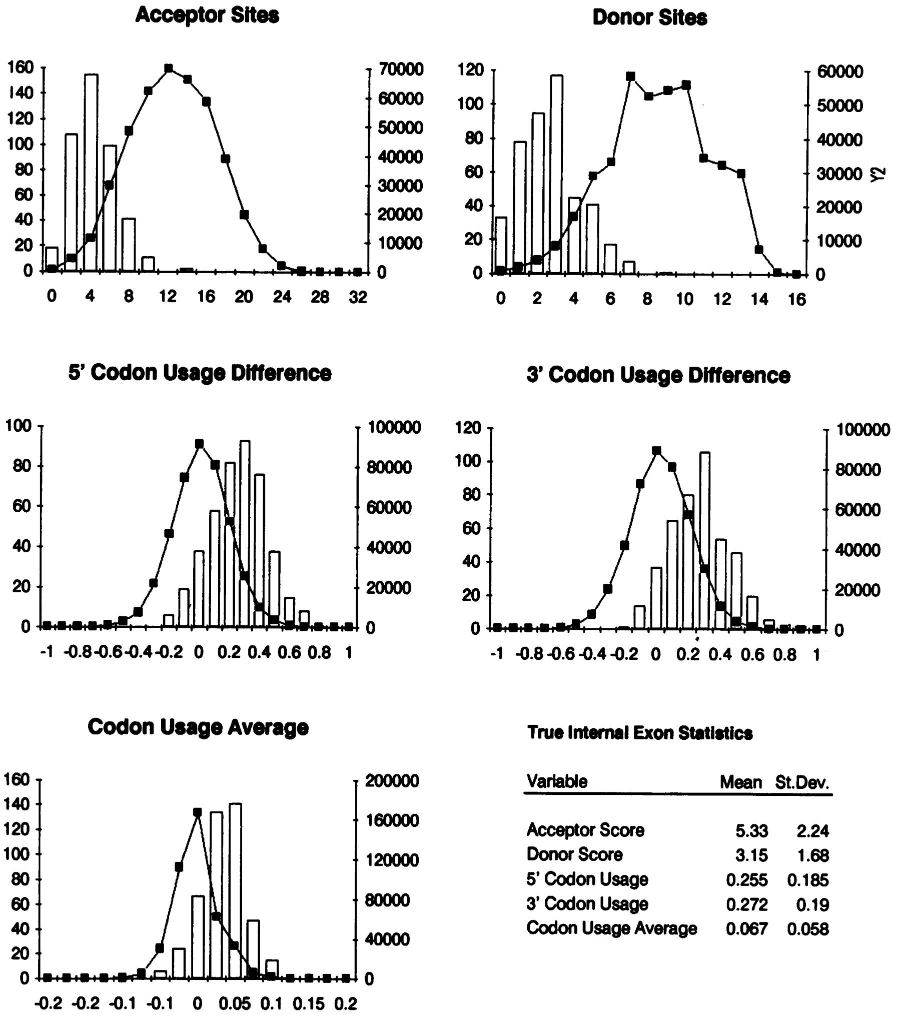
The output is a list of potential internal exons, ordered by confidence level, including start and stop positions, length, splice and codon usage scores, 5′ and 3′ phase and an amino acid translation. We define 5′ phase as the number of nucleotides required from a previous exon to put the current SORF into its correct reading frame. Similarly, we define 3′ phase as the number of nucleotides from the current SORF carried over to the next exon. In this way, adjacent SORFs can be compared to see if they fit together correctly. An incompatible fit suggests either that one of the predicted exons is in error, or an intervening exon is required to maintain the continuity of the reading frame. If the input is an annotated file in GenBank format, the output also includes a line for each SORF identifying its relationship to the exons described in the feature table. If internal exons identified in the feature table were missed, they are then displayed, giving reasons why the program failed to find them. This feature was added to assist the programmer in improving the algorithm, but in practice is also useful to suggest possible errors in splice site designation. In several instances, possible errors in GenBank annotations were discovered and were passed on to database editors.

The program was written in the C++ programming language on an IBM compatible 80286 computer. It has also been compiled for a Sun SparcStation running SunOs 4.1.2.

## RESULTS

### Determination of threshold parameters

In order to set the threshold rejection criteria for each of the variables calculated by the program, it was necessary to first determine the distribution of these variables for both true SORFS (those identical to known internal exons) and false SORFS. This was done for the 116 loci of the test set with the threshold parameters initially set to permissive values, with upstream AGs permitted in acceptor sites and with conflicts between SORFs allowed. As a result, the program identified 434 true internal exons out of 474 (91.6%), but also identified 470,828 false SORFS. Of the 40 true internal exons that were missed, 33 were less than 60 bp in length (7.0% of all internal exons), and the remainder were either lacking a donor GT consensus (5, or 1.1%)



**True Internal Exon Statistics**

| Variable | Mean | St.Dev. |
|---|---|---|
| Acceptor Score | 5.33 | 2.24 |
| Donor Score | 3.15 | 1.68 |
| 5' Codon Usage | 0.255 | 0.185 |
| 3' Codon Usage | 0.272 | 0.19 |
| Codon Usage Average | 0.067 | 0.058 |

**Figure 1.** Illustration of splice junction scores and codon usage statistics for true SORFs (those completely matching an internal exon) and false SORFS. In each case the number of true SORFs identified is shown as a bar graph and a line representing the number of false SORFs identified is overlaid. The ordinate for bar graphs (true SORFS) is on the left, whereas the ordinate for the line graphs (false SORFS) is on the right. The different scales were required due to the large excess of false SORFS. Note the difference in means in each case. The degree of difference in the means, and the variance of the distributions dictate the success of filtration at each step of the algorithm. The mean and standard deviation for the true SORFs are shown at lower right.

or were internal exons that preceded the coding region and contained stop codons (2, or 0.4%). The distribution of splice junction and codon usage scores for the true and false SORFs identified are shown in Figure 2. The different scales on the y-axis are necessary to demonstrate the distribution curves for both populations on the same graph, given the overwhelming predominance of false over true SORFS. Also note that while a given donor or acceptor site can appear in only one true SORF, identical sites may be shared by many false SORFs and so the number of sites shown on the graph exceeds the number of AG and GT dinucleotides in the sequences. Figure 2 illustrates why no one variable is capable of reliably discriminating between true and false predictions, and demonstrates that several filtration steps with set thresholds are necessary.

If the five variables were independent of one other, setting the threshold for each at its mean would lead to rejection of all but $(0.5)^5$, or 3.1% of the true SORFs and would also reject almost all of the false SORFS. Using this reasoning, the threshold values for confidence level 1 were set to the mean value for true exons in each of the distributions. For each succeeding confidence level, the thresholds were shifted by .464 standard deviations, so that by confidence level 5, approximately 97% of the true SORFs would be successful at passing any given threshold criterion. In practice, calculations based upon similar properties (such as the three related to codon usage), are not independent, and therefore a higher proportion of true SORFs survive each filtration step than that estimated here.

## Training set results

Table 3a presents an analysis of the training set (the 116 loci from which the distributions of variables were initially determined). The results are stratified by confidence level, showing the percent of the total internal exons greater than 60 bp in length (441) that are identified at each step. A complete match is defined as a SORF which shares precise splice junction boundaries with a known exon and is read in the same reading frame. Partial matches are those in which a SORF shares either the 5′ or 3′ splice junction (with or without the correct reading

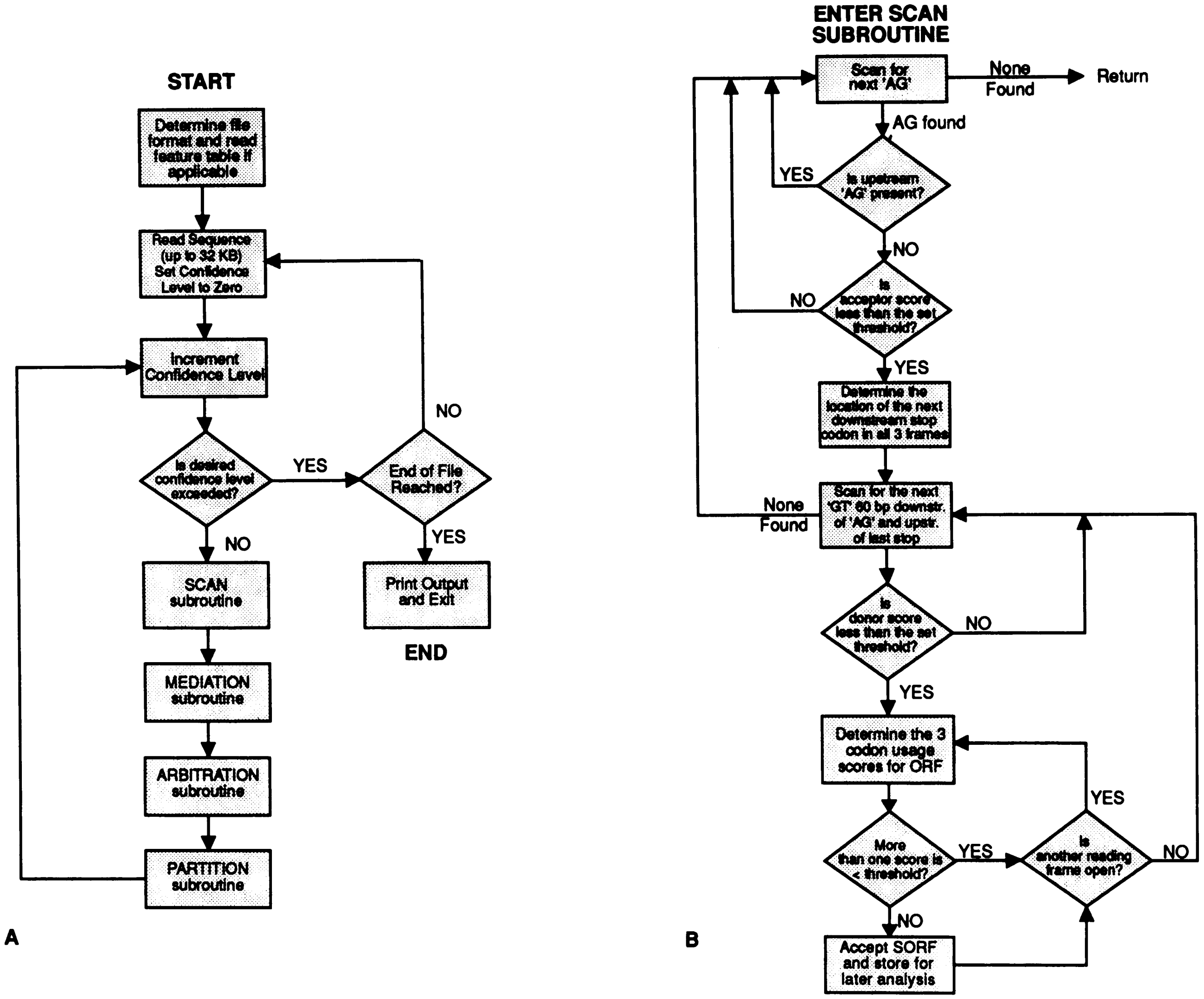

START
Determine file format and read feature table if applicable
Read Sequence (up to 32 KB) Set Confidence Level to Zero
Increment Confidence Level
Is desired confidence level exceeded?
YES
NO
End of File Reached?
NO
YES
SCAN subroutine
MEDIATION subroutine
ARBITRATION subroutine
PARTITION subroutine
Print Output and Exit
END
A
ENTER SCAN SUBROUTINE
Scan for next 'AG'
None Found
Return
AG found
Is upstream 'AG' present?
YES
NO
Is acceptor score less than the set threshold?
NO
YES
Determine the location of the next downstream stop codon in all 3 frames
Scan for the next 'GT' 60 bp downstr. of 'AG' and upstr. of last stop
None Found
Is donor score less than the set threshold?
NO
YES
Determine the 3 codon usage scores for ORF
More than one score is < threshold?
YES
Is another reading frame open?
YES
NO
NO
Accept SORF and store for later analysis
B

**Figure 2.** Algorithm Flowchart. Panel A depicts the main program. The SCAN subroutine is shown in more detail in panel B. The MEDIATION subroutine identifies SORFS that are mutually exclusive and the ARBITRATION subroutine chooses a single candidate exon for each mutually exclusive sequence segment. The PARTITION subroutine then marks those segments with successful SORFs so that SCAN does not search them again.

frame) , or both splice junctions but in the wrong reading frame. SORFs which are either completely separate from a true exon, or which overlap a true exon but do not share a splice junction are considered to be unmatched (false positives).

The overall sensitivity of the algorithm for identifying true internal exons of length greater than 60 bp is 87.1%. That is, by the time confidence level 5 has been reached, 280 (63.5%) of these true internal exons have been completely matched, and a further 104 (23.6%) have been partially matched. Different degrees of sensitivity are achieved at each confidence level. At confidence level 1, 9.3% of the true internal exons are completely matched. SORFs identified at this level of confidence are highly reliable; only 8.5% are false positives. However, at confidence level 5, where only 7.5% of true internal exons are identified, 84.2% of the SORFs are false positives. In practical terms, SORFs categorized in any of the first three confidence levels are more likely to correspond to a true exon than not, while SORFs at levels 4 and 5 are doubtful. If one includes all 5 confidence levels, 60.8% of SORFs identified to level 5 are false positives, with most of these false positive predictions occurring at the 4th or 5th confidence levels. The combined specificity of confidence levels 1 and 2, 1, 2 and 3 and 1, 2, 3 and 4 is 89.8%, 75.1% and 53.6%, respectively. It is noteworthy that 72 first exons (62.6%) and 38 last exons (32.8%) are partially matched, although the algorithm is not specifically designed to detect them since it requires splice junctions at each end of the open reading frame. In these cases, a sequence resembling a splice junction with a suitable score occurred by chance. Only 7% of the true internal exons in the training set are less than 60 bp in length. Including these exons in the calculations reduces the calculated sensitivity for all internal exons to 83.8%.

Partial matches to exons were further analyzed to determine the degree of overlap between the SORF identified and the actual internal exon. The 104 internal exons partially matched included 18010 nucleotides, of which 13949 were shared with the 18083 nucleotides of the identified SORFS, which represents a 77% overlap. Moreover, the predicted reading frame was found to be correct in 88.5% of the cases.

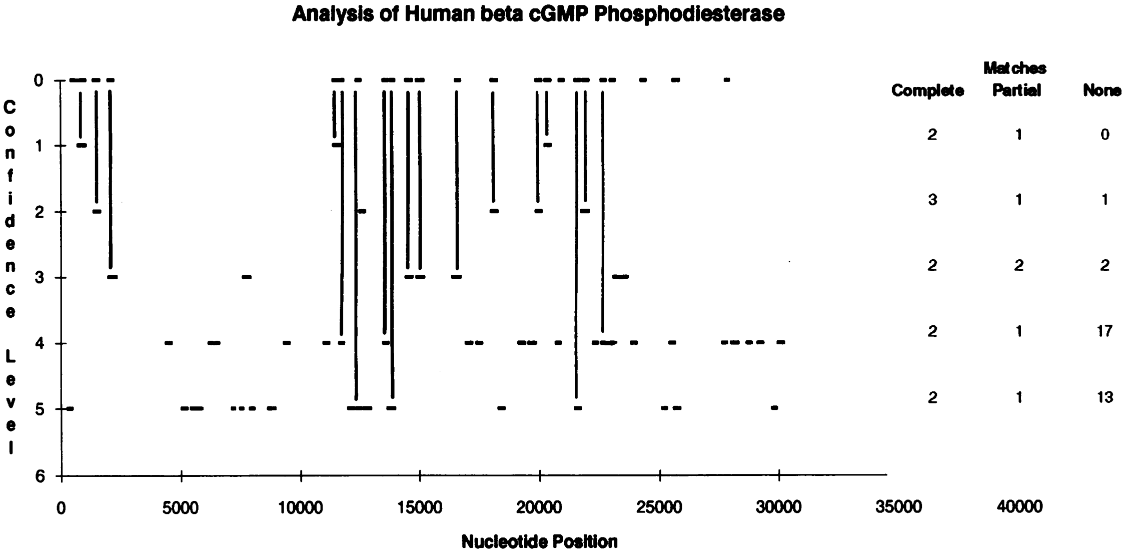


Figure 3. Results of an analysis of the beta-subunit of CGMP phosphodiesterase. The top line depicts the 22 exons of the gene, with one large intronic region between exons 3 and 4. SORFIND identifies 17 of the 22 exons, with 11 complete matches and 6 partial matches. The matches are depicted on the figure by vertical lines between the true exon and SORFs identified at levels of confidence shown on the Y-axis. The numbers at the right show the distribution of SORFs at each confidence level. Although there are 33 false SORFs overall, giving a specificity of 34%, 30 of these occur at confidence levels 4 and 5, making predictions at confidence levels 1 to 3 highly reliable.

## Test set results

An analysis of the 14 genes in the test set is also shown in Table 3b. These genes are independent of the 116 used to determine the thresholds for the algorithm, and thus provide a more rigorous test of the program. The results are not appreciably different from those of the training set, suggesting that the threshold levels calculated for the training set can be generalized to other genes of interest. Combined complete and partial matches identify 90.0% of the internal exons of length >60. Here 10.0% are found at confidence level 1, with 16.7% false positives. Specificity for the first 3 confidence levels is again greater than 50%, with overall specificity for all 5 levels of 34.5%.

The 23 SORFs that partially matched internal exons of the test set were also further analyzed to determine how well they represent the true exons. Of the 4272 nucleotides present in these SORFS, 4003 (93.7%) were shared with true internal exons. This accounted for 69.5% of the nucleotides in those exons. Of the 23 SORFs, 18 (78.3%) were read in the same reading frame as the true exon.

## Example of program usage

Figure 2 illustrates the results of an analysis of the beta-subunit of CGMP phosphodiesterase. This gene was sequenced in our laboratory (15), and is not part of the training or test data sets. In all, over 30 kb of genomic DNA from this gene has been sequenced. The gene consists of 22 exons, with one large intron between exons 3 and 4. The gene has been submitted to EMBL (accession numbers X62692–X62695) since four segmented entries as three intronic areas remain unsequenced. For the purposes of this example, the four segments have been joined together, separated by hyphens. SORFIND identifies 17 of the 22 exons, with 11 complete matches and 6 partial matches. Five of the six partial matches are interpreted in the correct reading frame. The matches are depicted on the figure by vertical lines between the true exon (top line) and SORFs identified at levels of confidence shown on the Y-axis. Although there are 33 false SORFs, giving an overall specificity of 34%, 30 of these occur at confidence levels 4 and 5. The predictions at confidence levels 1 to 3 highly are highly reliable, with 11 out of 13 SORFs (85%) corresponding to true exons.

## Comparison with other programs

Genomic sequence submitted to GenBank is generally biased because it rarely contains a significant quantity of flanking sequence. To truly assess the value of a program such as SORFIND, it should be used to analyze a large contig which contains several genes and a large amount of intergenic sequence. Moreover, the results should be compared with that of other existing software, such as CRM, the neural-net coding recognition module of the Gene Recognition and Analysis Internet Link (GRAIL) and GeneId, a hierarchical rule-based program which attempts to assemble entire genes. The sequence of two contigs (accession numbers M63796 and M89651) with 105,831 nucleotides spanning the ERCC1 locus of human chromosome 19q13.3 has recently been published (16), These two contigs contain, in addition to the ERCC1 gene, the human gene fosB, a third gene with partial homology to the rat type 2C protein phosphatase gene and two other expressed genes, A and B, with unknown function and with no homology to known genes. The authors included an analysis by CRM in conjunction with *gene modeler* (gm) to predict exon-intron structure in regions with positive CRM scores. Using the complete contigs, we examined the output of CRM, and compared it to that of SORFIND. We then compared the results of the GeneId program and SORFIND on 20 kb segments containing the ERCC1 and FOSB genes.

CRM selected ten open reading frames on the opposite strand of HUMMMDA which scored as 'Excellent', 'Good' or 'Marginal' in their potential as protein-coding regions. Four were associated with exons of the ERCC1 gene. Looking at the same sequence, SORFIND identified 14 candidate exons at confidence

**Table 3.** Results of the analysis of the training and testing sets. Numbers in brackets represent the percentage found of all internal exons greater than 60 bp in length. Specificity refers to the percentage of spliceable open reading frames (SORFs) identified that completely or partially match any exon. A partial match is defined as a SORF sharing at least one splice junction with a true exon, or sharing both splice junctions but interpreted in the wrong reading frame.

**Results on Training Set (116 genes, 441 internal exons >60 bp)**

| Confidence Level | 1 | 2 | 3 | 4 | 5 | Overall |
|---|---|---|---|---|---|---|
| Internal Exons >60 bp completely matched | 41( 9.3%) | 70(15.9%) | 82(18.6%) | 54(12.2%) | 33( 7.5%) | 280(63.5%) |
| Internal Exons >60 bp partially matched | 8( 1.8%) | 12( 2.7%) | 35( 7.9%) | 22( 5.0%) | 14( 3.2%) | 104(23.6%) |
| Total Internal Exons >60 bp matched | 49(11.1%) | 82(18.6%) | 117(26.5%) | 76(17.2%) | 47(10.7%) | 384(87.1%) |
| Internal Exons <60 bp partially matched | 0 | 2 | 2 | 6 | 3 | 13 |
| First Exons partially matched | 4 | 11 | 26 | 15 | 16 | 72(62.1%) |
| Last Exons partially matched | 1 | 5 | 7 | 13 | 12 | 38(32.8%) |
| Number of Sorfs unmatched to an exon | 5 | 14 | 87 | 268 | 416 | 790 |
| Total number of SORFs identified | 59 | 127 | 240 | 380 | 494 | 1300 |
| Specificity | 91.5% | 89.0% | 63.8% | 29.5% | 15.8% | 39.2% |

**Results on Test Set (14 genes, 80 internal exons >60 bp)**

| Confidence Level | 1 | 2 | 3 | 4 | 5 | Overall |
|---|---|---|---|---|---|---|
| Internal Exons >60 bp completely matched | 8(10.0%) | 13(16.3%) | 12(15.0%) | 9(11.3%) | 5( 6.3%) | 47(58.8%) |
| Internal Exons >60 bp partially matched | 2( 2.5%) | 5( 6.3%) | 8(10.0%) | 5( 6.3%) | 3( 3.8%) | 23(28.8%) |
| Total Internal Exons >60 bp matched | 10(12.5%) | 18(22.5%) | 20(25.0%) | 15(18.8%) | 9(11.3%) | 70(87.5%) |
| Internal Exons <60 bp partially matched | 0 | 0 | 0 | 1 | 1 | 2 |
| First Exons partially matched | 0 | 0 | 2 | 3 | 0 | 5(35.7%) |
| Last Exons partially matched | 0 | 0 | 3 | 0 | 0 | 3(21.4%) |
| Number of Sorfs unmatched to an exon | 2 | 3 | 14 | 52 | 85 | 156 |
| Total number of SORFs identified | 12 | 22 | 40 | 70 | 94 | 238 |
| Specificity | 83.3% | 86.4% | 65.0% | 25.7% | 9.6% | 34.5% |

**Table 4.** Comparison of SORFIND and CRM (GRAIL). The number preceding the slash indicates matches to true exons, whereas the number following the slash gives the number of candidates at that level of confidence (e.g. '2/4' implies that 2 candidates out of 4 matched true exons). SORFIND predicted exons in the regions of genes A and B(15) but since the exact boundaries of these genes have not been published, a question mark follows these numbers. CRM analyzes both strands whereas SORFIND requires a separate run with the reverse complement of the sequence. CRM does not give splice junction predictions, but requires the additional analysis of a program such as gm to do this. SORFIND considers only the best reading frame and provides an amino acid translation whereas CRM only suggests the best reading frame. 'CL' = confidence level.

| | | SORFIND | | | | CRM | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Contig** | **Gene(s)** | **CL1** | **CL2** | **CL3** | **Total** | **Excell.** | **Good** | **Marginal** | **Total** |
| HUMMMDA (pos. strand) 37314 bp | "A" | 0/0 | 1?/2 | 0/7 | 1?/9 | 1/2 | 0/2 | 0/0 | 1/4 |
| HUMMMDA (neg. strand) 37314 bp | ERCC1 | 2/2 | 3/5 | 1/7 | 6/14 | 2/2 | 2/3 | 0/5 | 4/10 |
| HUMMMDBC (pos. strand) 68505 bp | fosB & "phos" | 1/2 | 2/4 | 1/13 | 4/19 | 2/4 | 1/3 | 0/6 | 3/13 |
| HUMMMDBC (neg. strand) 68505 bp | "B" | 0/0 | 1?/3 | 1?/11 | 2?/14 | 2/4 | 0/5 | 0/6 | 2/15 |

**Table 5.** Comparison of SORFIND and GeneId. Only a 20 kb subsequence containing the entire gene was used in each comparison. The number of false positives is defined as the number of SORFs predicted by SORFIND or the number of exon equivalent classes predicted by GeneId that had the same or higher confidence level or score than the true predictions, but did not correspond to an exon. Each exon equivalent class predicted by GeneId can contain more than one predicted exon.

| Gene | Program | Exact match | Partial Match | Total number of Matches | Number of False Positives |
|---|---|---|---|---|---|
| ERCC1(10 exons) | GeneId | Exons 2,3,4,5,6,7, and 8 | Exon 9 | 8 | 17 |
| | SORFIND (to CL3) | Exons 3,4,5,6,7, and 9 | Exon 1 | 7 | 4 |
| | SORFIND (to CL4) | Exons 2,3,4,5,6,7,8 and 9 | Exon 1 | 9 | 10 |
| FOSB (4 exons) | GeneId | Exons 1,2,3 and 4 | | 4 | 11 |
| | SORFIND (to CL3) | Exon 2 | Exons 1 and 4 | 3 | 2 |
| | SORFIND (to CL4) | Exons 2 and 3 | Exons 1 and 4 | 4 | 14 |

levels 1, 2 or 3, of which 6 matched exons of ERCC1 exactly, with the correct amino acid translation. This predicted 219 out of the gene's 297 amino acids (74%) accurately. A seventh SORF correctly identified the 3′ junction of exon 1, but with the wrong reading frame. Two further exons were identified at confidence level 4, associated with 19 false positives. Two of fourteen open reading frames of HUMMMDBC selected by CRM corresponded to exons of the fosB gene and one contained a small fragment of the protein phosphatase gene. One exon of fosB was predicted exactly by SORFIND and another two partially, with 10 false positives. One SORFIND prediction matched 20 out of 39 (a 51% homology) of the amino acids of the rat protein phosphatase gene, suggesting that this represents a human exon of this gene. One SORF predicted in HUMMMDA is in the region of gene 'A' and two SORFs of the reverse complement of HUMMMDBC are in the region of gene 'B', but we do not know if these correspond to exons as the structure of these genes has not been published. These results are summarized in Table 4.

In order to undertake a comparison between SORFIND and GeneId, it was first necessary to truncate the input since GeneId can analyze a maximum of 20 kb. GeneId is designed to assemble entire genes, but it predicts individual exons at an earlier stage of analysis and categorizes its predictions into what are called 'equivalent exons'. With each predicted exon, the user is advised which reading frames are open. Each equivalent exon class may therefore contain several predicted segments, interpretable in one to three reading frames. GeneId determines the predicted reading frame when it assembles the exons into a complete gene. We compared SORFIND predictions with GeneId internal exon predictions (Table 5) rather than its final output. Two 20 kb subsequences containing the ERCC1 and fosb genes were used. GeneId predicted 7 ERCC1 exons with correct splice junctions, and predicted an eighth partially (with the correct 3′ splice junction). There were 17 incorrect 'equivalent exon' predictions. SORFIND identified 6 ERCC1 exons correctly (with the correct amino acid translation) in the first 3 confidence levels, There were 1 partially correct and 4 incorrect predictions. By extending the analysis to confidence level 4, SORFIND predicted all 8 ERCC1 internal exons correctly, with 1 partial match to the first exon and 10 incorrect matches. With fosB, GeneId predicted both internal exons, with 11 incorrect equivalent exon classes. SORFIND predicted 1 exon correctly at confidence level 3 (with 2 false positives), and the second at confidence level 4 (with a combined total of 14 false positives).

To summarize the comparisons, in the two contigs studied, SORFIND was more sensitive than CRM (at the expense of a higher number of false positive predictions) and similar in sensitivity to GeneId at predicting internal exons, but with higher specificity in the case of the ERCC1 gene.

## DISCUSSION

We have developed a program which will be a useful tool for screening naive genomic DNA sequence for regions likely to code for proteins, and which can be used in the laboratory using either an IBM-PC or SunOs Workstation. The program utilizes information on signal consensus, open reading frame and codon usage to predict exons given raw sequence as input. In a representative sample of genes from GenBank, it identifies 87% of the internal exons, with approximately 60% representing complete matches, and a further 27% being partially matched. Partially correct predictions share a large overlap with true exons, and are read in the correct reading frame over 80% of the time, making them useful in homology searches using the translated amino acid sequence, and permitting the design of specific probes for screening cDNA libraries. The specificity varies from 16% to 92%, depending upon the confidence level at which a prediction is made; with overall specificity from 35 to 40%. We suggest that researchers using SORFIND concentrate initially on those SORFs that are identified within the first three confidence levels, as our findings suggest that a specificity of greater than 60% can be achieved.

At this time, there exists little knowledge concerning pre-mRNA sequence which can be used to unambiguously predict how it will be spliced. The issue is complicated by the fact that pre-mRNA transcripts may be spliced in different ways at different stages of development in varying tissue types. Rule-based hierarchical algorithms, such as the one described here, rely upon filtration procedures that use a combination of thresholds which are set based upon the statistical properties of many known genes. As such, different properties which may direct splicing in a restricted tissue type or circumstance will be missed. Some variables, such as splice junction scores, are based upon properties that reflect the thermodynamics of DNA recognition by regulatory proteins (3, 4). Other properties, such as the presence of an open reading frame and codon usage bias, are artificial in the sense that they are unlikely to play a role in the actual mechanisms of RNA splicing. Nevertheless, variables based upon these artificial properties have a practical role to play in predicting splicing, while we await further elucidation of the true mechanisms involved.

Numerous factors may diminish the sensitivity of exon prediction programs during attempts to increase specificity. It is difficult, if not impossible, to design a procedure which will allow exceptions to the usual rules and which maintains a low number of false positive predictions. For example, a small but definite number of donor splice junctions have a non-standard consensus, with a GC rather than a GT dinucleotide just distal to the splice (17). An algorithm that would include these exceptions would need to filter a greatly increased number of potential exons, with a consequent significant reduction in specificity. A second example is the supposedly disallowed upstream AG in acceptor splice junctions (18), which nevertheless occurs in several instances in the training set. In addition, the codon usage statistics of particularly short exons do not display a variance that is narrow enough to allow the separation of true from false SORFS. Our program, SORFIND, effects a compromise by restricting its analysis to exons of at least 60 base pairs in length, with no near upstream AG dinucleotides in acceptor splice junctions, and no non-consensus donor splice junctions. In doing so it will inevitably miss a small number of true exons. The fact that sequencing errors do occur also reveals a problem with algorithms that rely upon open reading frames, for if an erroneous insertion or deletion introduces a frame-shift, altering the apparent codon usage or introducing a spurious nonsense codon, an exon will escape detection.

The arbitration procedure utilized by the program serves to reduce the number of false positive SORFS, but also on occasion eliminates a true exon from consideration when a false SORF successfully competes with it. These instances, where a false SORF may appear in all respects superior to the exon that is actually translated, highlight the current incomplete knowledge of pre-mRNA splicing mechanisms.

Some computer programs for exon prediction maximize sensitivity by setting liberal thresholds which will allow passage of all but a few true exons through each stage in the filtration procedure. The successful filtration of true from false exons then depends upon the number of variables examined, and the degree of independence they have from one another. Our program differs in that it initially sets very strict thresholds in order to isolate those SORFs of high confidence; it then gradually relaxes the thresholds, reducing specificity as the sensitivity increases. This serves to focus attention on those predictions with highest probability, while still identifying as many potential exons as possible. This approach has allowed identification of close to 90% of true internal exons in a training and test set of genes. Furthermore, SORFIND identified 17 of 22 (77%) exons of a gene identified in our laboratory that was not part of the test or training sets. As such, this program may be useful for those wishing to identify coding sequence in long stretches of genomic DNA. Translations of these DNA sequences into their amino acid residues, provided by the program, may also improve database searches for homologous proteins.

SORFIND is comparable to other currently used programs that predict protein-coding regions, but any such comparison is necessarily approximate, as each program provides output in different formats and is intended to accomplish different tasks. In the two contigs we used for comparison, CRM appeared to be less sensitive (exons missed) but more specific (fewer false positives). CRM may perform better when there are several sequencing errors that interrupt open reading frames. GeneId is designed to assemble entire genes, and so is currently more complete in its prediction of first and last exons, but in this analysis gave a greater number of false predictions with similar sensitivity. SORFIND has an advantage over both programs in providing an amino acid prediction which is correct in the great majority of cases, and which can be used immediately as input to a protein homology search program such as BLAST. It is thus likely to be useful to identify genes when only incomplete sequence is available.

## AVAILABILITY

The binary version of the program, which will run on any IBM-compatible microcomputer with an 80286 microprocessor, or better, is freely available. There is an identical version which runs under SunOs. Details may be obtained by writing the authors at the above address, or by e-mail on internet to hutch@ulam.generes.ca.

## ACKNOWLEDGMENTS

The constructive suggestions and encouragement of Dr Allen Delaney during the development of this program have been greatly appreciated. We also wish to thank Dr Bernhard Weber and Colin Collins for their helpful comments on the manuscript. This research is supported by grants from MRC-Canada and the Canadian Genetic Diseases Network. M.R.H. is an established investigator of the Children's Hospital of BC.

## REFERENCES

1. Staden,R. (1990) In Doolittle,R. (ed.), *Methods in Enzymology*. **183**, 163−180.
2. Stormo,G.D. (1990) In Doolittle,R. (ed.), *Methods in Enzymology*. **183**, 211−221.
3. Berg,O.G. and von Hippel,P.H. (1987) *J. Mol. Biol.* **193**, 723−750.
4. Berg,O.G. and von Hippel,P.H. (1988) *J. Mol. Biol.* **200**, 709−723.
5. Lapedes,A., Barnes,C., Burks,C., Farber,R. and Sirotkin,K. (1988) In Bell,G. and Marr,T. (eds), *Computers and DNA*. Addison-Wesley. pp. 157−182.
6. Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) *J Mol. Biol.* **220**, 49−65.
7. Claverie,J., Sauvaget,I. and Bougueleret,L. (1990) In Doolittle,R. (ed.), *Methods in Enzymology*. **183**. 237−252.
8. Konopka,A. and Owens,J. (1990) *Gene Anal. Techn Appl.* **7**, 35−38.
9. Uberbacher,E.C. and Mural,R.J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 11261−11265.
10. Fields,C. and Soderlund,C. (1990) *CABIOS* **6**, 263−270.
11. FickettJ.W. (1982) *Nucleic Acids Res.* **10**, 5303−5318.
12. Gelfand,M.S. (1990) *Nucleic Acids Res.* **18**, 5865−5869.
13. Guigo,R., Knudsen,S., Drake,N. and Smith,T. (1992) *J. Mol. Biol.* in press.
14. Penotti,F.E. (1991) *J. Theor. Biol.* **150**, 385−420.
15. Weber,B., Riess,O., Hutchinson,G., Collins,C., Lin,B., Kowbel,D., Andrew,S., Schappert,K. and Hayden,M.R. (1991) *Nucleic Acids Res.* **19**, 6263−6268.
16. Martin-Gallardo,A, McCombie,W.R., Gocayne,J.D., FitzGerald,M.G., Wallace,S., Lee,B.M.B., Lamerdin,J., Trapp,S, Kelley,J.M., Liu,L.-I., Dubnick,M, Johnston-Dow,L.A., Kerlavage,A.R., de Jong,P., Carrano,A., Fields,C. and Venter,C. (1992) *Nature Genetics* **1**, 34−39.
17. Senapathy,P., Shapiro,M.B. and Harris,N.L. (1990) In Doolittle,R. (ed.), *Methods in Enzymology*. **183**, 252−278.
18. Mount,S.M. (1982) *Nucleic Acids Res.* **20**, 459−472.