Nucleic Acids Research, 2003, Vol. 31, No. 22 ***e141***
DOI: 10.1093/nar/gng141

FELINES: a utility for extracting and examining EST-defined introns and exons

Scott D. Drabenstot, Doris M. Kupfer[1], James D. White[1], David W. Dyer, Bruce A. Roe[1], Kent L. Buchanan[2] and Juneann W. Murphy*

Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, PO Box 26901, BMSB 1053, Oklahoma City, OK 73190, USA, [1]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK 73019, USA and [2]Department of Microbiology and Immunology, Tulane University Medical Center, New Orleans, LA 70112, USA

Received July 31, 2003; Revised and Accepted September 23, 2003

ABSTRACT

FELINES (Finding and Examining Lots of Intron 'N' Exon Sequences) is a utility written to automate construction and analysis of high quality intron and exon sequence databases produced from EST (expressed sequence tag) to genomic sequence alignments. We demonstrated the various programs of the FELINES utility by creating intron and exon sequence databases for the fungal organism *Schizosaccharomyces pombe* from alignments of EST to genomic sequences. In addition, we analyzed our constructed *S.pombe* sequence databases and the well-established *Saccharomyces cerevisiae* intron database from Manuel Ares' Laboratory for conserved sequence motifs. FELINES was shown to be useful for characterizing branchsites, polypyrimidine tracts and 5′ and 3′ splice sites in the intron databases and exonic splicing enhancers (ESEs) in *S.pombe* exons. FELINES is available at http://www.genome.ou.edu/informatics.html.

INTRODUCTION

Expressed sequence tags (ESTs) provide a snapshot of gene expression in an organism (1) and are valuable for a variety of techniques including microarray construction and development of gene prediction software for genome annotation. As the genomes of more eukaryotic organisms are sequenced, the value of EST libraries is increasingly recognized for defining introns and exons in genomic sequences. For example, the Genomic On-Line Database (GOLD) lists over 270 ongoing or complete eukaryotic sequencing projects (2). The GenBank EST database, dbEST (http://www.ncbi.nlm.nih.gov/dbEST/index.html), lists 559 different organisms with ESTs; 345 organisms have 100 or more ESTs (3). Automated analysis processes are necessary to take advantage of the enormous number of EST and genomic sequences available in the public domain. Although several programs can align ESTs to genomic sequences (4–6), none of these tools allows for the creation of high quality intron and exon sequence databases derived from alignments. Also, public domain utilities for mining such databases are limited.

Our objective was to develop a utility to assemble intron and exon sequence databases based upon EST to genomic alignments and analyze those databases for conserved elements. Rather than attempt to create a new EST to genomic sequence alignment program, an existing program was used. After reviewing available alignment programs including est2genome (5), Spidey (6) and sim4 (4), Spidey was selected for reasons of speed and sensitivity (6). The utility FELINES (Finding and Examining Lots of Intron 'N' Exon Sequences) that we developed and describe here provides a means of automated construction and analyses of intron and exon sequence databases. This utility operates at three levels. The first level is the alignment process that pairs ESTs with their homologous genomic sequences and aligns the ESTs on the genomic sequence using BLASTN and Spidey in a batchwise automated manner. The second level filters and extracts intron and exon sequences from the EST to genomic sequence alignments. The third level allows the user to analyze the constructed intron and exon sequence databases.

To validate and demonstrate the utility and versatility of FELINES, we used sequence data from two model fungal organisms. The first organism was the fission yeast, *Schizosaccharomyces pombe*. *Schizosaccharomyces pombe* is an ideal test organism because the genomic sequencing has been completed and a large number of EST sequences are available. In addition, *S.pombe* genes often contain one or more introns (7). These features enabled us to demonstrate the entire FELINES utility. The second organism was the budding yeast, *Saccharomyces cerevisiae. Saccharomyces cerevisiae* was chosen as a test organism for the FELINES utility due to the availability of an annotated database of all known *S.cerevisiae* introns (8). Thus, the *S.cerevisiae* dataset allowed us to validate aspects of FELINES. In addition, the *S.cerevisiae* intron dataset served as an excellent example of the analysis process for an externally derived intron sequence database using the FELINES utility.

*To whom correspondence should be addressed. Tel: +1 405 271 6622; Fax: +1 405 271 3117; Email: juneann-murphy@ouhsc.edu

Table 1. List of functional ESE motifs used to identify ESE elements in *S.pombe* exons (26)

Protein bound	Functional ESE motif: Perl regular expression	Functional ESE motif: IUPAC expression
SRp20	GCUCCUCUUCC	GCUCCUCUUCC
	CCUCGUCC	CCUCGUCC
SC35	G[AG][CU][CU][AC]C[CU][AG]	GRYYMCYR
SF2/ASF	C[AG][CG][AC][CG]G[AU]	CRSMSGW
SRp40	[CU][AG]C[AG][GU][AC]	YRCRKM
SRp55	[CU][CU][AU]C[AU][GC]G	YYWCWSG

From Cartegni *et al.* (26). Brackets indicate that either nucleotide is possible.

MATERIALS AND METHODS

Definitions

The nucleic acid abbreviations used were based upon the standard IUPAC abbreviations (9). Briefly, A = adenine; C = cytosine; G = guanine; U = uridine; Y = cytosine or uridine; R = guanine or adenine; W = adenine or uridine; S = guanine or cytosine; K = guanine or thymine; N = uridine, cytosine, guanine or adenine.

Motif definitions

The FELINES utility is designed to allow customization of parameters that are organism dependent or that the investigator wishes to define more or less stringently. For example, one program of the FELINES utility automatically searches for motifs such as branchsites or polypyrimidine tracts; the user can customize the definition for these motifs within the FELINES options file. Each motif is defined using Perl regular expression which enables complex motifs to be defined (10). For all conserved sequences other than the branchsite, the motifs were identified using a linear search process.

The motifs that we used for searching our *S.pombe* and *S.cerevisiae* datasets for branchsites, polypyrimidine tracts and possible ESEs are listed below. The branchsite motif definition used was based on branchsite motifs previously described for metazoans and *S.cerevisiae* (8,11,12). Briefly, the primary branchsite motif was defined as CURAY, the secondary branchsite was defined as UURAY, and the alternative was a modified YURAY motif where either the 1st, 3rd, or 5th position could be an N while the other positions were held constant. The modified YURAY motif was used only if no CURAY motif or UURAY motif could be identified. FELINES identified the potential branchsite as the last instance of the motif in the 3′ end of the intron. The polypyrimidine tract definition we used was based upon the results of previous studies of polypyrimidine tracts (13–16). Briefly, a polypyrimidine tract was defined as any six consecutive nucleotides that contained at least three uridines and no adenines. The functional human ESE motifs used for searching for potential ESEs in the *S.pombe* exon dataset are listed in Table 1.

Data sources and data preparation

To demonstrate the utility of FELINES, a database of 8123 *S.pombe* ESTs was downloaded from dbEST on January 15, 2003. The genomic sequences used were *S.pombe* chromosomes 1, 2 and 3. These sequences were downloaded from GenBank on July 10, 2003 and formatted for BLAST analysis using formatdb (17). A separate FASTA file was created for each EST sequence and genomic sequence using all2many.pl (J.White and B.Roe, unpublished, University of Oklahoma, http://www.genome.ou.edu/informatics.html). The Ares Laboratory Yeast Intron Database (YIBD) 3.0 (8) is a database of 253 introns of *S.cerevisiae* representing 100% of that organism's known introns with the branchsite for each intron identified. We obtained this dataset from http://www.cse.ucsc.edu/research/compbio/yeast_introns.html and used it without modification.

Additional program sources

The Linux executable of Spidey (6) was obtained from the National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nih.gov/pub/wheelan/Spidey/) and used without modification. The Linux executable of blastall (17) was also obtained from NCBI (ftp://ftp.ncbi.nih.gov/pub/blast/executables/) and used without modification. Perl version 5.8.0 was obtained from http://www.perl.org/. Some programs in the FELINES package use the Perl module Statistics::Lite, available from the Comprehensive Perl Archive Network (CPAN, http://www.cpan.org).

RESULTS

Schema for identification of introns, exons and conserved motifs

Figure 1 shows the manner in which the FELINES utility can be used to construct intron and exon sequence databases and to search for conserved motifs contained in the intron and exon datasets. The arrows illustrate the path of data flow. The process of using FELINES begins by customizing the options file which is drawn on directly by other FELINES segments. The Alignment Level or Level 1 is next, and in this step wiscrs.pl is run to provide the Spidey alignments for the EST and genomic sequence pairs employing the user-specified conditions in the options file. Extraction Level or Level 2 follows with gumbie.pl parsing the Spidey alignment files based upon the conditions specified in the options file. The output of gumbie.pl provides intron and exon sequence databases. It is at this point that externally prepared intron or exon datasets can be inserted. The final step is the Analysis Level or Level 3 in which three different programs, namely icat.pl, cattracts.pl and findnmers.pl, are available to define conserved motifs such as branchsites, polypyrimidine tracts, 5′ and 3′ splice junction motifs or potential ESEs.

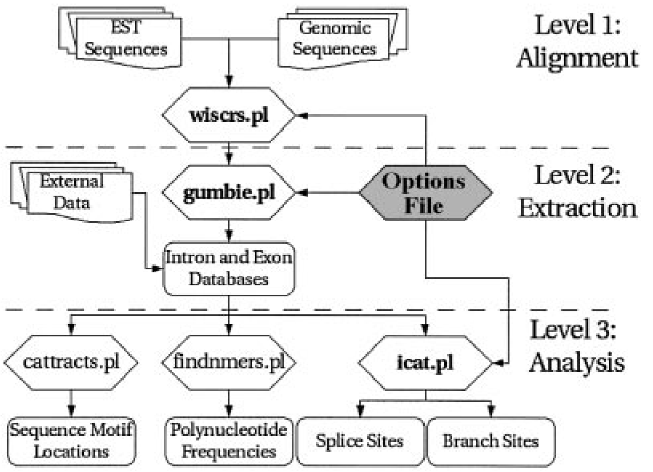


Figure 1. Data flow showing the use of the programs in the FELINES utility.

FELINES components

FELINES options file. Because many of the features of introns and exons vary between organisms, FELINES was designed to allow for user customization. The FELINES options file allows the user to define the options that will be used by multiple FELINES components. This simplified customization increases the integration of the utility because the multiple programs can reference a single options file. If one desires to search the databases using varying parameter sets, options files can be designed for each set of parameters one desires to use. This feature enables concurrent comparisons of results from two or more different options settings. For example, when choosing the e-value for the initial BLASTN analysis performed by wiscrs.pl, it is possible to see how using a smaller or larger e-value affects the size of the resulting intron or exon sequence databases. For our *S.pombe* example, we established a customized options file that can be found in the supplementary data (http://www.genome.ou.edu/informatics.html).. The specific options we used for *S.pombe* and the logic in selecting those options are discussed below with each FELINES component that draws from the options file.

Automated batchwise alignments—wiscrs.pl. As noted above, the initial goal was to align large numbers of EST sequences to their corresponding genomic sequences and use these alignments to create intron and exon sequence databases. The wiscrs.pl (write individual sequence comparisons, run spidey) program was written to automate the alignment layer in this process. The first step of this process was to pair EST sequences with related genomic sequences. Though there are many different methods for determining sequence similarity, BLAST is perhaps the most familiar and is widely available (17). Therefore, using BLASTN, wiscrs.pl created an output file of homologous ESTs for each genomic sequence. Whether or not an EST is determined to be a potential homolog to a genomic sequence depends to a great extent on how 'homology' is quantified. One of the most widely used methods for determining homology is the e-value derived through a BLAST comparison. A user can decide what e-value constitutes homology and set that e-value in the FELINES options file as the minimal acceptable e-value. We found that more restrictive e-values reduce the computational time but can also result in valid alignments being excluded from the resulting intron and exon databases. For our *S.pombe* dataset example, an e-value of 1 × 10^{-3} was set in the options file. Wiscrs.pl performed the BLASTN comparison, and then created a file containing all EST sequences deemed homologous for each genomic sequence.

The second step that wiscrs.pl does is initiation of EST to genomic sequence alignments by Spidey. This results in one extended file for each genomic sequence containing the best possible alignment of each EST to the genomic sequence.

Creation of intron and exon databases—gumbie.pl. After the ESTs have been successfully aligned to the genomic sequences, the next step was to extract the genomic alignment coordinates from the Spidey alignment files and create intron and exon sequence databases from the alignments. Gumbie.pl (gather up mRNA-based introns and exons) was written for this purpose. Gumbie.pl performs the extraction layer of the overall process (Fig. 1) and was based, in part, on extractfasta.pl (J.White and B.Roe, unpublished, University of Oklahoma, http://www.genome.ou.edu/informatics.html). At the command line, gumbie.pl is given the name of the options file and the name of the file containing the names of all of the Spidey alignment files. Gumbie.pl extracts the identified intron and exon regions from the genomic sequence and parses them to the respective database.

The sequence databases were constructed in either raw or filtered mode. In raw mode, gumbie.pl did not filter any of the alignments. However, because (i) a single EST may have short regions of homology to multiple genomic regions, (ii) the genomic or EST sequences may contain sequencing errors, or (iii) small exons may be incorrectly aligned (6), it was desirable to filter the sequences before constructing the databases. To eliminate these problems, multiple, easily modified filtering criteria were built into FELINES. Each filtering criterion was designed to be fully customizable from within the options file.

The first set of alignment filters evaluates global alignment quality. Global alignment filters are effective at removing weak alignments quickly. The filters include minimum percent identity, minimum percent mRNA coverage, maximum number of allowable mismatches, maximum number of allowable gaps and the minimum number of exons. The three most powerful filters were the minimum percent identity, the minimum percent coverage and the minimum number of exons. These three filters effectively screened out the majority of the poor alignments.

The complete options file that we used for our *S.pombe* example is available in the supplementary data.

The minimum percent identity is a measure of the number of identical nucleotides between two sequences over a defined sequence distance. In development of the Spidey program, Wheelan *et al.* (6) required a 95% or greater similarity. However, the data used in the development of Spidey consisted of mRNA annotations extracted from GenBank. EST sequences, as single pass sequences, are more error prone than mRNA annotations from GenBank. Consequently, the majority of mRNA annotations in GenBank match the genomic sequence exactly, whereas EST sequences often have regions of low quality sequences that may have poor homology to the corresponding genomic sequence. Therefore, when analyzing the alignment of the *S.pombe* EST sequences to the genomic sequence, a cut-off level of 90% sequence identity was chosen and was the parameter placed in the *S.pombe* options file.

The minimum percent coverage is a measure of the length of the mRNA covered by the sequence alignment. To better accommodate the problems of low quality sequences found in EST libraries, Spidey is designed to truncate the EST sequences in order to enable alignment of the EST to the genomic sequence. However, allowing too much truncation results in a meaningless alignment. In our *S.pombe* test set, 80% minimum coverage was used.

The remaining three global alignment filters, concerned with maximum number of allowable mismatches, maximum number of allowable gaps and the minimum number of exons, were used to remove alignments that were generally weak but still met the above criteria. In this way, sequences that appeared to have a significant number of gaps in the alignment or to have a great number of mismatches can be removed.

The second set of filters used with gumbie.pl contained local alignment filters. These allowed for the evaluation of the quality of individual intron and exon sequences. Local alignment filters were very effective in removing intron and exon sequences that had poor quality alignment only over a part of the EST (i.e. short exons not accurately aligned by Spidey or poor quality sequence data). These filters examined intron length, splice site pairs and intronless exon length.

Intron length was also a useful local alignment filter. In the process of aligning ESTs to genomic sequences, Spidey is designed to align exons at all costs. This occasionally resulted in introns that were tens-of-thousands of nucleotides in length, a situation that is quite unrealistic for introns of the fungal organisms we were analyzing. Eukaryotes are known to efficiently splice introns that fall in a defined length range, and that length range varies between species. For example, Deutsch and Long (18) found that the intron length for humans ranged from 25 to 54 916 nt and for *Aspergillus* from 42 to 241 nt. So, we wanted to set parameters into the options file that would eliminate introns that were unrealistically long for eukaryotic organisms like fungi. To determine the parameters to use in the options file, we reviewed typical intron lengths for other eukaryotic organisms. We found that the shortest eukaryotic spliceosomal introns found to date were identified in the chlorarachniophyte algal endosymbiont, and had an average length of 20 nucleotides (19). Thus, the length of 20 nucleotides was put into the options file as the shortest allowed intron length for *S.pombe*. We used 2000 nucleotides as the maximum acceptable intron length for *S.pombe*, because this is about twice the size of the longest identified intron in the fungus *S.cerevisiae* (20). Therefore, the acceptable intron-length range in the options file for *S.pombe* was set to 20–2000 nucleotides.

A second local alignment filter contained the definition for an acceptable intron splice site pair. It has been reported that 5′GU...AG3′ and 5′GC...AG3′ account for more than 98% of the annotated introns in GenBank (18,21). In addition, recently published whole genome sequences suggest that 5′GU...AG3′ and 5′GC...AG3′ intron splice sites are the only splice classes used by *S.pombe* or the related hemiascomycetous fungi (7,8,22). As a result, only 5′GY...AG3′ splice sites were considered valid in this study. It is important to note, however, these are not the only splice site classes that have been identified (23). Consequently, in the FELINES options file, one can define as valid any desired splice class.

Exon sequences have considerably greater variation in length and lack strongly conserved 5′ or 3′ splice site sequence motifs. Therefore, construction of useful exon sequence filters was based upon the quality of flanking introns. As an example, if an intron with two flanking exons did not meet the defined filtering criteria, neither of the flanking exons would be included in the exon sequence database.

There were also a number of exons that did not have flanking introns. These 'intronless exons' obviously could not be filtered based upon the quality of the adjacent introns. However, we desired to include intronless exons in the exon sequence database as they presumably represent a gene with an uninterrupted open reading frame. As a result, a minimum length filter was implemented for intronless exons. For analyzing the *S.pombe* data, the minimum length used was 300 nt. This length is consistent with the minimum length often used for definition of ORFs (24). As a result of using the above filters, a database of 1298 intron sequences and a database of 4024 exon sequences were created from 8123 EST sequences and the three *S.pombe* chromosomal sequences. An example of a portion of the gumbie.pl output is available in the supplementary data.

Identifying conserved intron motifs—icat.pl. Icat.pl (intron consensus analysis tool) was developed to perform two tasks. First, icat.pl was designed to filter all intron databases generated outside of FELINES using the parameters from the options file. This insured that all intron datasets generated outside of the FELINES utility were comparable to those generated from within the utility. A part of the icat.pl summary output describes the number of sequences removed from any intron dataset analyzed and the basis for the removal (see supplementary data). The second task for which icat.pl was developed was identification of conserved sequence motifs in intron sequences. The conserved sequence motifs for which the program searched were defined in the FELINES options file and the results of the search were reported in the icat.pl summary output (see supplementary data).

When we ran our *S.pombe* intron sequence dataset through icat.pl, there were no additional introns removed. This was expected because icat.pl and gumbie.pl used the same options file, and we did not make any changes to that options file before we ran icat.pl. We also ran the Ares Laboratory *S.cerevisiae* YIDB (8) through icat.pl using an options file identical to the one we customized for *S.pombe*. No introns from the YIDB dataset were removed by icat.pl. Below we have described the use of icat.pl for finding conserved sequence motifs such as branchsites, polypyrimidine tracts, and 5′ and 3′ splice sites.

Branchsite identification. The algorithm used to identify the branchsite in icat.pl was designed to search for three variations of the branchsite motif, the primary, the secondary and an alternate motif. To identify the potential branchsite sequence, all overlapping instances of the primary motif, the secondary motif and the alternate branchsite motif were identified within each intron sequence. If the primary motif was found, the 3′-most instance of the primary motif was considered the branchsite. If there were no instances of the primary motif, the 3′-most instance of the secondary motif was considered the branchsite. Finally, if neither the primary nor secondary motifs were found, the 3′ most instance of the alternative motif was identified as the potential branchsite. Any introns for which a potential branchsite sequence could not be identified were not included in subsequent branchsite-related analyses.

Using the information about the potential branchsite, icat.pl went on to create four additional files in FASTA format. Each of these files only contained introns for which a potential branchsite sequence could be identified. The first file was a list of the complete sequence for each intron. The second file contained the branchsite sequence for each intron. The third file contained the sequence of each intron 5′ of the branchpoint, i.e. the sequences from the 5′ splice site through the branchpoint 'A'. The fourth file contained the sequence of the intron after the potential branchpoint, i.e. the sequences beginning after the branchpoint 'A' to the 3′ end of the intron.

To validate icat.pl, we compared the annotated branchsites for the YIDB with the results obtained by running the YIDB through icat.pl. The branchsite motifs used were CURAY as the primary motif, UURAY as the secondary motif and YURAY (allowing the variations noted in Materials and Methods) as the alternate motif. The YIDB is an extensively tested database of intron sequences with annotated branchsites for *S.cerevisiae*, the majority of which have been validated *in vivo* or *in vitro* (8). By checking the branchsite identified by icat.pl for congruence with the YIDB annotated branchsite, we verified that icat.pl was able to find branchsites effectively. Both the YIDB and the icat.pl program identified a branchsite for 100% of the introns in the YIDB dataset. We found that for 90% of the introns, the branchsite identified by icat.pl was identical to the annotated branchsite in the YIDB. For most introns where icat.pl identified an alternate branchsite from the YIDB, there were two, or in one case three, overlapping CUAAC motifs, i.e. CUAACUAAC or CUAACUAACU-AAC. Based upon the above algorithm, icat.pl identified the CUAAC motif closest to the 3′ end of the intron as the branchsite. In the YIDB, the branchsite was annotated as the CUAAC motif closest to the 5′ end of the intron. Using the same primary, secondary and alternate motifs with our *S.pombe* intron dataset, we were able to identify a branchsite motif for 99.9% of the introns.

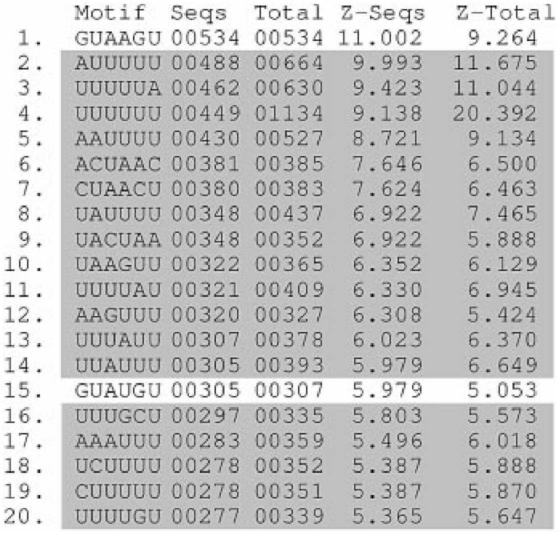
Polypyrimidine tract identification. Icat.pl was used to identify all the relative locations of polypyrimidine tracts in the *S.pombe* and *S.cerevisiae* introns based upon the definition for polypyrimidine tracts that we had placed in the FELINES options file. We found potential polypyrimidine tracts in 90% of the *S.pombe* introns surveyed and 97% of the *S.cerevisiae* introns surveyed. The introns for which no polypyrimidine tract could be identified ranged in length between 35 and 105 nucleotides.

Icat.pl also separately searched for polypyrimidine tracts in the region of the intron sequence from the 5′ splice site to the branchpoint 'A' and the region from the branchpoint 'A' to the 3′ splice site. Each intron was classified as having (i) polypyrimidine tract(s) 5′ to the branchpoint only, (ii) polypyrimidine tract(s) 3′ to the branchpoint only, (iii) polypyrimidine tract(s) 5′ and 3′ to the branchpoint, or (iv) no polypyrimidine tracts that met the options file definition. We found that while 29% of the *S.cerevisiae* introns analyzed had polypyrimidine tracts only in the 5′ region of the intron, 62% of the *S.pombe* introns had polypyrimidine tracts only between the 5′ splice site and the branchpoint. Conversely, we identified a polypyrimidine tract only in the region from the branchpoint to the 3′ splice site in 3.6% of the *S.cerevisiae* introns and 1.6% of the *S.pombe* introns.

5′ and 3′ splice sites. As previously mentioned, icat.pl is also useful for identifying the 5′ and 3′ terminal nucleotides. Using icat.pl, it was also noted that the 5′GU...AG3′ terminal dinucleotides of the intron accounted for all but one (>99%) of the *S.pombe* introns in our database. In the YIDB intron dataset, 98% of the introns matched the 5′GU...AG3′ terminal dinucleotide motif. The remainder of the introns surveyed had the 5′GC...AG3′ splice site motif. A previous whole genome analysis of *S.pombe* intron splice sites based on a predicted intron dataset found three 5′GC...AG3′ introns (7).

Identifying fixed-length motifs—findnmers.pl. Findnmers.pl was designed to identify all fixed-length polynucleotide sequences (N-mers) in a single FASTA file. For example, findnmers.pl could be used to find all pentamers, hexamers, decamers, etc. in a group of sequences. Any over-represented motifs could be identified and could represent conserved motifs in the sequences analyzed. Findnmers.pl was designed to allow the user to specify the length of the N-mer at the command line. Findnmers.pl then tallies the number of occurrences of each N-mer. In addition, a Z-score is calculated for each N-mer. The Z-score and other statistics are approximated assuming a uniform DNA base distribution of N-mers with equal N-mer probabilities. The occurrence of any one N-mer can be modeled as a binomial with probability, $p = 1/4^N$, $q = 1 - p = (4N - 1)/4^N$. A further approximation is to use a Poisson distribution using a parameter equal to the number of N-mers examined times p. These distributions, however, are only an approximation, because the DNA base distribution is not truly random, and the N-mers will not be independent of each other when sliding through a sequence one base at a time as we did in findnmers.pl. Both approximations were used in findnmers.pl, consequently, results from findnmers.pl provide only an estimate of the Z-scores for the most abundantly represented N-mers. As a command line option, the user can specify a minimum Z-score for an N-mer to be included in the output. Z-score was included to provide a means of determining how over-represented an N-mer was.

We analyzed our *S.pombe* intron database using the findnmers.pl default N-mer length of six nucleotides to discover any over-represented nucleotide hexamers. Findnmers.pl tallied the number of sequences in which the hexamer occurred (Seqs), and the total number of occurrences of each hexamer (Total). Findnmers.pl also calculated Z-scores for the number of sequences in which each hexamer occurred (Z-Seqs) and the total number of occurrences of each hexamer (Z-Total). A partial output for our *S.pombe* intron database is shown in Figure 2. In the *S.pombe* dataset, the first hexamer, GUAAGU was the most common six-nucleotide



```
      Motif  Seqs  Total Z-Seqs  Z-Total
 1.   GUAAGU 00534 00534 11.002   9.264
 2.   AUUUUU 00488 00664  9.993  11.675
 3.   UUUUUA 00462 00630  9.423  11.044
 4.   UUUUUU 00449 01134  9.138  20.392
 5.   AAUUUU 00430 00527  8.721   9.134
 6.   ACUAAC 00381 00385  7.646   6.500
 7.   CUAACU 00380 00383  7.624   6.463
 8.   UAUUUU 00348 00437  6.922   7.465
 9.   UACUAA 00348 00352  6.922   5.888
10.   UAAGUU 00322 00365  6.352   6.129
11.   UUUUAU 00321 00409  6.330   6.945
12.   AAGUUU 00320 00327  6.308   5.424
13.   UUUAUU 00307 00378  6.023   6.370
14.   UUAUUU 00305 00393  5.979   6.649
15.   GUAUGU 00305 00307  5.979   5.053
16.   UUUGCU 00297 00335  5.803   5.573
17.   AAAUUU 00283 00359  5.496   6.018
18.   UCUUUU 00278 00352  5.387   5.888
19.   CUUUUU 00278 00351  5.387   5.870
20.   UUUUGU 00277 00339  5.365   5.647
```

Figure 2. Partial output from findnmers.pl for our *S.pombe* intron database.

motif detected by findnmers.pl, and it occurred once in each of 534 introns. Since this sequence was strongly reminiscent of the 5′ terminal intron motif, we decided to look for similar sequences. We found the 15th hexamer, GUAUGU, occurred a total of 307 times in 305 different introns (Fig. 2). The 6th, 7th and 9th hexamers appear to be branchsite motifs, each containing at least a portion of the CUAAC motif commonly identified as the branchsite (Fig. 2). The 10th, 12th and 17th hexamers are not as easily categorized (Fig. 2). However, the 10th and 12th hexamers (Fig. 2) matched a repeat motif that has been categorized for zebra fish (accession no. AL590146) and the 17th hexamer matched a repeat motif identified for humans (accession no. AL031846). The remaining hexamers are likely portions of polypyrimidine tracts based upon their high uridine content. An extended portion of the output from this analysis can be found in the supplementary data. Having identified potential conserved sequence motifs, we were interested in discovering where in the introns these motifs were located. For this purpose, we used cattracts.pl, described below.

Locating user-defined sequence motifs. Cattracts.pl can search for multiple user-defined sequence motifs simultaneously. The input for cattracts.pl is the name of the FASTA-formatted sequence file to be searched and the name of a FASTA-formatted flat file containing the desired motif(s) written in Perl regular expression format (10).

We first wished to confirm the polypyrimidine tract findings from icat.pl. A search for polypyrimidine tracts using the definition described in the Materials and Methods, provided identical results to those identified with icat.pl.

Next, we decided to use cattracts.pl to extend the findings from findnmers.pl. We searched for the two previously identified hexamers, GUAUGU and GUAAGU using cattracts.pl. The hexamer GUAAGU occurred in the first 10 nucleotides of 533 introns. The single remaining instance of the motif was in the 71–80 nucleotide range. These data indicated that these sequences are indeed present nearly exclusively at the 5′ end of *S.pombe* introns. The hexamer GUAUGU was found in the 0–10 nucleotide range in 299 instances, 21–30 nucleotide range in three instances, twice in the 101–110 range and once in the 121–130 nucleotide range. Cattracts.pl also found that there were four introns with two occurrences of GUAUGU or GUAAGU while the remainder of the introns containing either of the two hexamers had only one occurrence each. A portion of this cattracts.pl analysis is available in the supplementary data.

We also used cattracts.pl to search the exon sequences for conserved sequence elements. Some of the most common conserved exon sequences are exonic splicing enhancers (ESEs). ESEs are important for the definition of exons and have been found to regulate alternative splicing by binding serine-arginine (SR) proteins (25). We searched the exon regions of our *S.pombe* exon sequence database using a list of consensus sequences for functional ESEs (Table 1) (26). Of the 3944 exons searched, only 32 had regions that matched the SRp20 functional ESE motif while 2479 had regions that matched the SRp40 functional motif and 1491 had regions that matched the SRp55 functional motif. This is interesting given that the already identified *S.pombe* SR protein homolog, Srp2, is most similar to the human SR proteins SRp40 and SRp55 (27). In addition, we found that 3159 or 80.1% of the exons surveyed had motifs that matched putative ESE motifs. Our findings using cattracts.pl to identify potential ESE motifs confirm and extend the findings of Lutzelberger *et al.* (27).

DISCUSSION

We have demonstrated that the FELINES package is a useful tool kit for constructing high quality datasets of introns and exons and for manipulating the datasets in a fast, automated, modular manner. Due to the rapidly increasing size and number of EST and genomic libraries, FELINES is predicted to be a highly useful utility for accurately defining introns and exons. Wiscrs.pl in conjunction with gumbie.pl makes the process of obtaining high quality intron and exon datasets a task that can be easily performed on a PC computer attached to the internet. We used wiscrs.pl and gumbie.pl to create *S.pombe* intron and exon databases derived from ESTs aligned to genomic sequences. The intron and exon databases we created were significantly larger than previous experimentally based *S.pombe* intron and exon datasets, and were of high quality because of the filtering process incorporated into gumbie.pl. The quality of the datasets generated through the use of wiscrs.pl and gumbie.pl can be varied to suit the investigator simply by modifying parameters in the options file. In fact, one can set up several options files containing different numbers or definitions for each parameter, so that one can determine how changing any given parameter will affect the data outcome. FELINES does not limit one to starting the process with EST and genomic data for preparing intron and exon datasets. As we demonstrated with the *S.cerevisiae* intron dataset that was already available from the Ares Laboratory, one can analyze data in existing datasets using the Analysis level of FELINES (Fig. 1).

Before using the Analysis level of FELINES to obtain information from our *S.pombe* datasets, we wanted to validate the utility and quality of the programs using a fungal intron dataset from an organism with an annotated database of all of its known genes. The *S.cerevisiae* intron dataset of Ares

(YIBD) fits this need and was subjected to icat.pl. Icat.pl was designed to identify conserved sequence motifs in intron sequences and to filter existing intron datasets using the same criteria as used by gumbie.pl for introns. Icat.pl filters did not eliminate any of the *S.cerevisiae* introns, which was an expected result, based on the quality of the database. Our search for branchsites showed that icat.pl identified potential branchsites in all *S.cerevisiae* intron sequences and, for 90% of the introns, the branchsites were identical to branchsites previous defined (8). The majority of the miscalls were minor in that overlapping branchsite motifs were present and icat.pl called the motif closest to the 3′ end of the intron and the YIDB-defined branchsites were the motifs closest to the 5′ end. The excellent agreement of our branchsite results with those annotated in the YIDB confirmed that icat.pl could accurately and rapidly identify nucleotide motifs in a database.

Using icat.pl, we were able to identify polypyrimidine tracts and show the relationship of the polypyrimidine tract location to the location of the branchpoint of the intron. Icat.pl was also useful for confirming that the 5′GU...AG3′ splice site class was the predominant intron donor/acceptor pair used in *S.pombe*. Finally, icat.pl enabled the identification of 5′GC...AG3′ splice class introns and our findings agreed with previous findings of Wood *et al.* (7).

The FELINES program findnmers.pl enabled us to examine over-represented intron sequence motifs of any length we desired. Using findnmers.pl to search for hexamers, we identified two potential 5′ splice site motifs. In addition, other frequently occurring hexamers were suggestive of motifs previously defined in other organisms, or the hexamers were potentially polypyrimidine tracts. The splice site motifs and polypyrimidine tracts identified are important binding sites in spliceosomal formation (28), but they are not the only conserved motifs that might show up in a search for over-represented sequences. Identification of over-represented sequences has been an effective means of identifying other conserved sequence motifs (29,30).

Another analysis component of FELINES, cattracts.pl was used to provide data to confirm that the putative 5′ splice site motifs that were suggested by findnmers.pl output were indeed 5′ splice site motifs. Cattracts.pl was also extremely useful in identifying motifs that could possibly be ESEs in exonic sequences of *S.pombe*. Using cattracts.pl with the *S.pombe* exon database, we determined that over 80% of the exons surveyed contained a sequence region that matched one of the previously identified functional human ESE motifs.

Due to the interrelated nature of the information retrieved using each Analysis level program, the findings from one program served to confirm the findings of another program. For instance, the polypyrimidine tract results were identical in icat.pl and in cattracts.pl. Furthermore, the program findnmers.pl identified uridine-rich sequences as over-represented sequences. Also, when the 5′ branchsite motifs were analyzed using icat.pl, cattracts.pl and findnmers.pl, each program identified the 5′GU motif as the predominant splice motif.

The FELINES programs can be used to compare branchsite, polypyrimidine tracts, 5′ and 3′ splice sites, and putative ESE elements from datasets from several organisms. This application was demonstrated by using intron and exon databases prepared from EST and genomic sequences from several fungi, namely the Basidiomycota, *Cryptococcus neoformans* and two Ascomycota, *Aspergillus nidulans* and *Neurospora crassa* (D.M.Kupfer, S.D.Drabenstot, K.L.Buchanan, H.Lai, H.Zhu, D.W.Dyer, B.A.Roe and J.W.Murphy, submitted for publication). We compared branchsite sequences, polypyrimidine tracts, 5′ and 3′ splice sites, and putative ESE occurrences of each organism with those found in our *S.pombe* intron and exon sequence databases (D.M.Kupfer, S.D.Drabenstot, K.L.Buchanan, H.Lai, H.Zhu, D.W.Dyer, B.A.Roe and J.W.Murphy, submitted for publication).

FELINES is a versatile, fast, automated modular utility. FELINES is quickly and easily adapted for use with a wide variety of eukaryotic organisms by modifying the options file. The FELINES utility can analyze small or large sequence datasets; there is no theoretical limitation to the size of the datasets that can be analyzed. These strengths make FELINES a very useful utility for rapid and efficient analysis of multiple genomes. Furthermore, although FELINES was written on a Linux system, it is written in Perl and can be used on virtually any platform.

Availability

The FELINES package can be downloaded from the World Wide Web at http://www.genome.ou.edu/informatics.html. Users of this package should cite the present publication as a reference. The authors welcome comments, corrections and additions.

ACKNOWLEDGEMENTS

Special thanks to Hongshing Lai for help in producing the first version of gumbie.pl, and to Josh Orvis who wrote the predecessor to findnmers.pl. This work was supported by a grant from the National Institutes of Health (AI-47079).

REFERENCES

1. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
2. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes Online Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.*, **29**, 126–127.
3. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—Database for 'Expressed Sequence Tags'. *Nature Genet.*, **4**, 332–333.
4. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
5. Mott,R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS*, **13**, 477–478.
6. Wheelan,S.J., Church,D.M. and Ostell,J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952–1957.
7. Wood,V., Gwilliam,R., Rajandream,M.-A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
8. Grate,L. and Ares,J.M. (2002) Searching yeast intron data at Ares lab Web site. *Methods Enzymol.*, **350**, 380–392.
9. IUPAC-IUB Commission on Biochemical Nomenclature (CBN) (1971) Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations. *Arch. Biochem. Biophys.*, **145**, 425–436.
10. Wall,L., Christiansen,T. and Orwant,J. (2000) *Programming Perl*. 3rd edn. O'Reilly and Associates, Cambridge, MA.
11. Burge,C.B., Tuschl,T. and Sharp,P.A. (1999) In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World—The Nature of Modern*

RNA Suggests a Prebiotic RNA. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 525–560.

12. Berglund,J.A., Chua,K., Abovich,N., Reed,R. and Rosbash,M. (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, **89**, 781–787.
13. Coolidge,C.J., Seely,R.J. and Patton,J.G. (1997) Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.*, **25**, 888–896.
14. Roscigno,R.F., Weiner,M. and Garcia-Blanco,M.A. (1993) A mutational analysis of the polypyrimidine tract of introns. *J. Biol. Chem.*, **268**, 11222–11229.
15. Shelley,C.S. and Baralle,F.E. (1987) Deletion analysis of a unique 3′ splice site indicates that alternating guanine and thymine residues represent an efficient splicing signal. *Nucleic Acids Res.*, **15**, 3787–3799.
16. Singh,R., Valcarcel,J. and Green,M.R. (1995) Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science*, **268**, 1173–1176.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Deutsch,M. and Long,M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, **27**, 3219–3228.
19. Gilson,P.R. and McFadden,G.I. (1996) The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed and the smallest known spliceosomal introns. *Proc. Natl Acad. Sci. USA*, **93**, 7737–7742.
20. Spingola,M., Grate,L., Haussler,D. and Ares,J.M. (1999) Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA*, **5**, 221–234.
21. Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the exon-intron database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
22. Bon,E., Casaregola,S., Blandin,G., Llorente,B., Neuveglise,C., Munsterkotter,M., Guldener,U., Mewes,H., Helden,J.V., Dujon,B. *et al.* (2003) Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.*, **31**, 1121–1135.
23. Mount,S.M. (2000) Genome sequence, splicing and gene annotation. *Am. J. Hum. Genet.*, **67**, 788–792.
24. Furuno,M., Kasukawa,T., Saito,R., Adachi,J., Suzuki,H., Baldarelli,R., Hayashizaki,Y. and Okazaki,Y. (2003) CDS annotation in full-length cDNA sequence. *Genome Res.*, **13**, 1478–1487.
25. Graveley,B.R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6**, 1197–1211.
26. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.
27. Lutzelberger,M., Gross,T. and Kaufer,N.F. (1999) Srp2, an SR protein family member of fission yeast: *in vivo* characterization of its modular domains. *Nucleic Acids Res.*, **27**, 2618–2626.
28. Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
29. Shah,A., Giddings,M., Parvaz,J., Gesteland,R.F., Atkins,J.F. and Ivanov,I. (2002) Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, **18**, 1046–1053.
30. Hahn,M., Stajich,J. and Wray,G. (2003) The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.*, **20**, 901–906.