[22] Searching Yeast Intron Data at Ares Lab Web Site

By LESLIE GRATE and MANUEL ARES, JR.

Introduction

It must be obvious to every geneticist by now that the future will be consumed by the need to understand how the elemental properties of genes so elegantly described in the past half-century come together with the environment to produce the subtle differences that are key to the fitness of the organism. This will require a partial abandonment of the reductionism so favored since Mendel, to be replaced by the adoption of a more synthetic view that addresses the molecular underpinnings of complex phenotypes, penetrance, expressivity, and the small contributions of many genes. Although many of us were trained to design experiments about single genes, or at the most two interacting genes, our students and researchers need more. More in this case is a healthy computational philosophy and experience.

We have tried to embrace this in our own small way by setting up a searchable database containing information concerning the introns found in the genome of *Saccharomyces cerevisiae*. Since one of us (MA) has training in genes but not computers, and the other (LG) has training in computers but not genes, this effort has been a cultural compromise. Despite its lack of sophistication and dotcom sheen, the database has found many uses in our laboratory and has been accessed by yeast geneticists, splicers, and bioinformaticists the world around. In the following pages we explain the browsing and search capabilities of the site, and how to read and interpret the findings.

Getting Into the Site

Probably the best way to use this chapter is to sit at the computer with the book open, as you go through the descriptions of the different searches. Although some readers may find computers intimidating, there is really no way to damage equipment or files using a Web browser, so no big mistakes can be made. Explore! Experiment! The site can be found by following the "Ares Lab Yeast Intron Database" link from the Ares lab home page at http://ribonode.ucsc.edu. The site can also be accessed from the *Saccharomyces* Genome Database (SGD[1]) by clicking on their "Yeast WWW Sites" link and scrolling down to their "Yeast Introns" link. Alternatively, access the site directly by typing "http://www.cse.ucsc.edu/research/compbio/yeast introns.html" into the location window of your favorite browser, and hit return.

Au: Pls. supply all author's names. Series style not to use et al. in reference list.

[1] C. A. Ball *et al.*, *Nucleic Acids Res.* **28,** 77 (2000).

0076-6879/02 $35.00

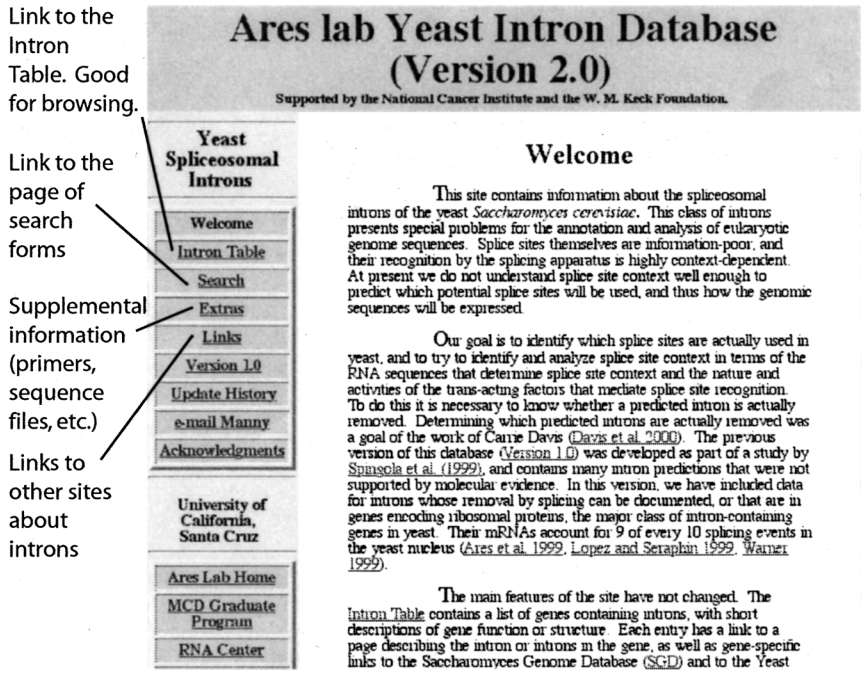


FIG. 1. Front page of the Ares lab Yeast Intron Database. **Au: Cap l?**

The front page is shown in Fig. 1. In addition to some text about introns and where the information comes from, this page has four important links, two of which we will discuss at length. The "Intron Table" link (Fig. 1) will load a large document that includes a table that has the yeast genes with introns (Fig. 2). The next link on the navigation bar, "Search" links to a page (Fig. 3) that contains a set of links to different types of searches that can be performed (discussed below). The "Extras" link goes to a page that contains additional information of interest, such as available PCR primers for detecting splicing of individual introns, text files of intron sequences and alignments, and various graphs and other data related to introns in yeast. The "Links" page includes links to other Web sites concerning introns, of note, Seraphin's yeast site[2] and Kent's *Caenorhabditis elegans* Intronerator.[3] Click on the "Intron Table" link and look at the Intron Table page.

Browsing the Intron Table

The Intron Table page is shown in Fig. 2. Each entry has a link to an "Ares Lab Intron Report" which is represented by the entry number. The entries in the table are

[2] P. J. Lopez and B. Seraphin, *Nucleic Acids Res.* **28,** 85 (2000).
[3] W. J. Kent and A. M. Zahler, *Nucleic Acids Res.* **28,** 91 (2000).

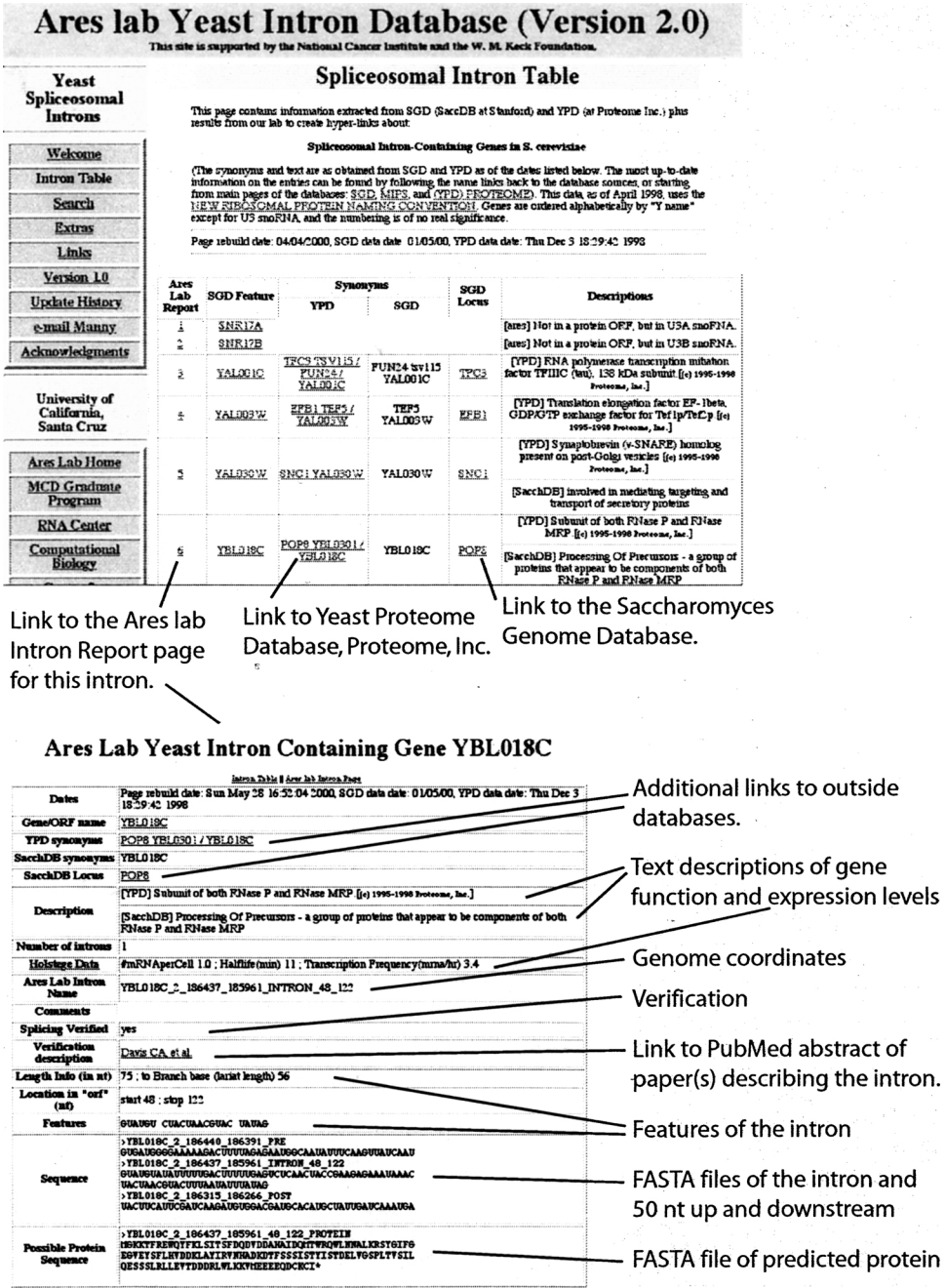


FIG. 2. The Intron Table and the Intron Report pages.

listed in alphabetical order by ORF name or "Y name" (e.g., YAL001C, meaning: Y, yeast; A, first chromosome; L, left arm; three digit number, assigned to ORF; C, the Crick strand), with the exception of the snoRNA genes U3 genes *SNR17A* and *SNR17B*. These are listed first, RNA coming before protein. Each entry contains a link to the SGD locus page for that gene, as well as information concerning synonyms, gene names (as opposed to ORF names), and a short description of the function of the gene product.

Clicking on the entry number link in the column "Ares lab report" produces a page for the gene in that entry, shown in the lower part of Fig. 2. This report shows much of the data in the underlying database for the intron-containing gene specified at the top of the page. The first four rows of the report reiterate the information and links on the Intron Table, including the short text descriptions. Additional information such as number of introns, expression level of the gene in the experiment by Holstege *et al.,*[4] the genome coordinates of the intron, any comments we had, and whether or not splicing has been verified experimentally or is predicted are included in the next rows. A link to the Genbank file (if one is available), or the PubMed abstract of the paper describing molecular evidence for splicing, or the method of prediction is included next. Next, physical features of the intron are presented. Length, position relative to the AUG of the ORF (except for 5′ UTR introns), and the starting, ending, and branchpoint region sequences of the intron are listed.

In the row labeled Sequence, there are three FASTA files (a "FASTA file" is a standard file format for presenting sequence data) associated with each intron. These include sequences 50 nucleotides (nt) upstream of the 5′ splice site (ending with the label "PRE"), the sequence of the intron itself, and the first 50 nt following the intron (labeled "POST"). These sequences can be copied for use with other kinds of sequence searches and alignment programs. Be aware that precise transcription initiation sites for most yeast genes are unknown, and our arbitrary use of 50 nt upstream of the intron does not imply that initiation occurs more than 50 nt upstream of the intron in every case. An example would be the U3 genes in which the first exon is only 16 nucleotides.[5] The last line, "Possible Protein Sequence," shows a protein prediction from sequence that has had the intron removed. It is accurate for genes with a single intron in the coding sequence, but the program we use for this may generate faulty predictions for genes with two introns or with introns in the 5′ UTR. Therefore, the protein predictions should be used with care.

The Intron Table page is a large (∼200 K) document, and clicking back and forth from it to the report pages can be slow, even with a fast connection. The best way to avoid reloading this large page is to open it once, and open a new

[4] F. C. Holstege *et al., Cell* **95,** 717 (1998). **Au: Please supply all names.**

[5] E. Myslinski, V. Segault, and C. Branlant, *Science* **247,** 1213 (1990).

browser window in which to view the report. Usually the right button on a three-button mouse will open a link in a new window. With a single-button mouse, hold down the button with the cursor over the link until the pop-up menu appears on the screen near the cursor. Select "New window with this link" or "Open link in new window" and release the mouse button. The linked report page should open in a new window that is displayed over the top of the Intron Table. When done viewing the report, close the new window and return to the Intron Table window. This avoids having to reload the Intron Table page each time. A little bit of practice will make these operations go more smoothly.

Other features of most browsers are also useful, for example, the "Find . . ." function available under the "Edit" menu. Selecting "Find . . ." opens a small dialog box with a space in which to type what you want to find in the open page. We use this to find things in the Intron Table without scrolling up and down forever. Try it by searching for the actin intron: type "actin" in the dialog box and clicking "Find." The browser searches the text for "actin" and highlights the first occurrence of this string of letters. If we start at the top of the Intron Table, the first occurrence of "actin" is in the *ARP2* gene entry (actin-related protein). Using "Find again" to go to the next occurrence, we find SAC6 (actin filament bundling protein), and finally we come to *ACT1* itself. Other intron-containing genes that work with actin can be found by continuing the process to the end of the Table. This approach is good for browsing for key words as well as for finding a particular gene in the Table by name.

Searching the Intron Database

The simple searches offered by the browser are limited to short text strings (i.e., sequences of letters and numbers) and only in the page displayed. More complex queries can be generated using four types of searches available by clicking "Search" on the navigation bar on the left of most pages from the site. The Search page is shown in Fig. 3 and basically contains links to each of the searches. The first is the Intron Table Text Query, which allows selected properties of introns to be defined and returns a smaller table identical in format to the large Intron Table, but which only contains the entries that match the query. The Intron Splice Signals page can be used to identify introns that have particular branchpoint or splice site sequences of interest. The YAG Query page allows specifics concerning the 3′ splice site and the region between the branchpoint and the 3′ splice site to be captured. Finally, the Intron Sequence Search allows identification of introns containing a particular nucleotide sequence of interest. Below we will describe examples of how to use these search functions.

Go to the Intron Database Text search page (Fig. 4) from the "Search" page by clicking on the "Intron Table Text Query" link. This page provides a more sophisticated means to find and organize a subset of introns of interest. One can search

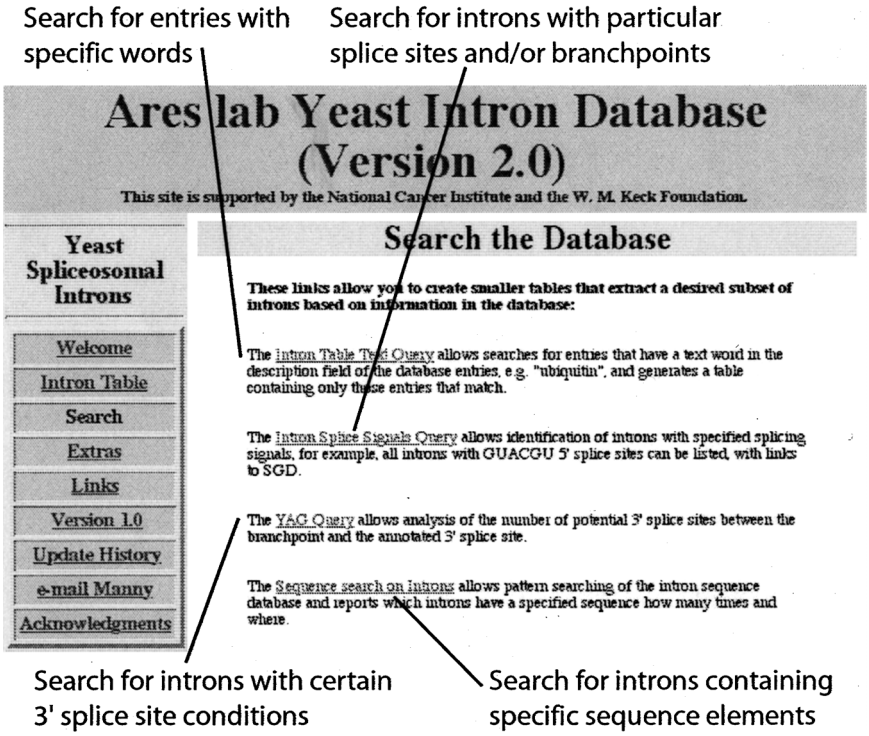


FIG. 3. The Search Page.

for introns by different properties, such as description text words, gene names, comments, intron number, and the length of different parts of the intron itself. The example described in Fig. 4 shows the output from a search for "alternative splicing" in the comments box. Submitting this request (by typing "alternative splicing" and clicking on the "Submit intron table query" button) returns a page that has a small table identical in form to the large Intron Table, but which only contains the entries that fit the terms requested. In this case there are two yeast genes known to have alternative splicing, *MTR2* and *SRC1*.[6] Each of these appears in a row as it would in the large Table, but without the other introns. As with the large Intron Table, clicking on the entry number generates the intron report page for that entry. This type of search allows rapid winnowing of lists of introns to obtain those that fit certain criteria of interest.

Multiple constraints can be imposed on the search, because the form treats each piece of information submitted in the different boxes as "AND" rather than

[6] C. A. Davis *et al.*, *Nucleic Acids Res.* **28,** 1700 (2000).

Au: Please supply all names.

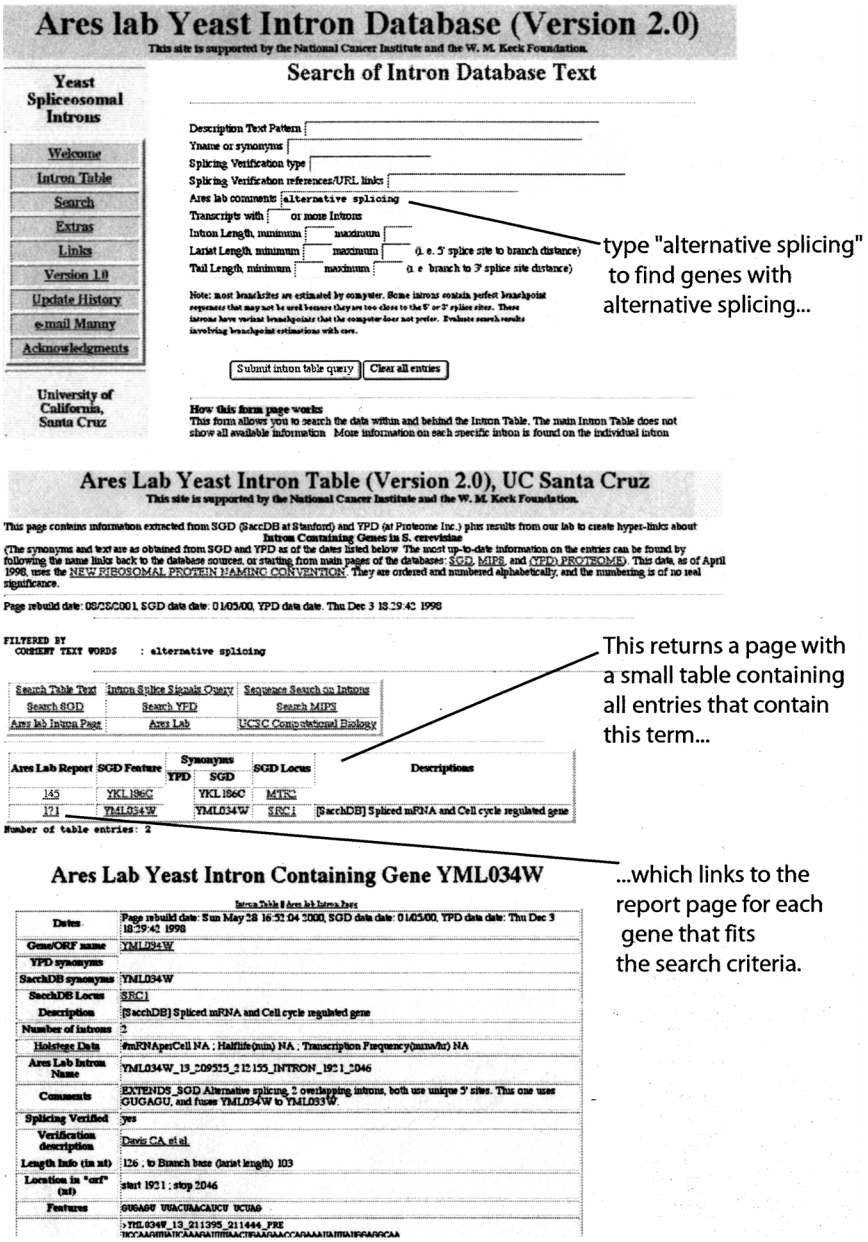


FIG. 4. Text searches to obtain subsets of genes of interest.

"OR." For example, try typing "UTR" in the comments box and submitting the form. This will return all genes with introns in the 5′ UTR. Note the number of introns. Now try typing "ribosomal protein" in the "Description Text Pattern" box and "UTR" in the comments box. This will only return ribosomal protein genes that have introns in their 5′ UTRs. This is a search that retrieves too many results, but it can be made more restrictive. To make a search less restrictive, one can use a partial word as a text pattern (since a "text pattern" is not exactly the same as a word). For example, the text pattern "ubi" recalls genes encoding ubiquitin, ubiquitin-conjugating enzymes, and ubiquinol cytochrome *c* reductase subunits. This can sometimes have unintended effects. Try typing "actin" in the "Description Text Pattern" box and clicking on the "Submit intron table query" button. Most of the results returned will be of interest; however, note the presence of *YIP2* on the list. Why is this gene here? It is not similar to actin or known to be involved in actin function. The computer found the sequence "actin" in the text description "Ypt interacting protein." This shows that humans will always remain important in evaluating results from computational processes, and that each result should be considered innocent until proven guilty.

An intron feature commonly of interest is the sequence of the splice sites and branchpoint. The set of yeast introns has highly conserved splicing signals[7]; thus deviations are of interest. Go back to the "Search" page (Fig. 4) and then click on the "Intron Splice Signals Query" link. The form page for the Splice Site Query is shown in Fig. 5. To find all introns with the nonstandard branchpoint GACUAAC (rather than the most common UACUAAC sequence) type "gacua" into the "Bpre" box on the form. The form will appear with "a" in the Branch box, since all yeast introns are thought to use A as the branched nucleotide. Leaving the other boxes open is the same as asking for anything at those positions. Click on the "Submit intron query" button. As shown in Fig. 5, the search identifies 10 introns that contain GACUAAC as the best match to the consensus. Each intron is listed with its intron signals and a link to its report page. Having the other splicing signals displayed shows, for example, that *YGL251C* and *YLR211C* also have unusual 5′ splice sites, and that *YDL115C* has an unusual 3′ splice site. The search can be extended to reveal splice site and branchpoint context as well. For example, to search for all introns containing four U residues immediately upstream of a GACUAAC branchpoint, one would type "uuuugacua" into the "Bpre" box and submit the form. The five introns that fit this description would then be returned. The other boxes work in the same way, allowing the intronic context of splice signals to be explored.

We became interested in what the intron collection might tell us about 3′ splice site selection in yeast, and we developed a search that allows identification of the number of potential 3′ splice site–like sequences that might exist between the

[7] M. Spingola *et al., RNA* **5,** 221 (1999).

Au: Please supply all authors' names.

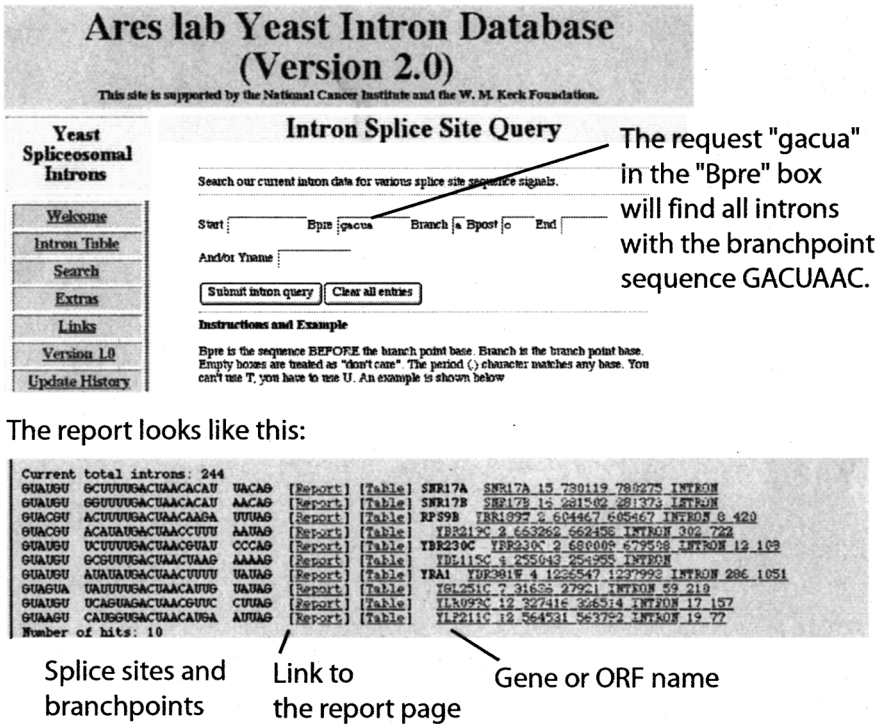
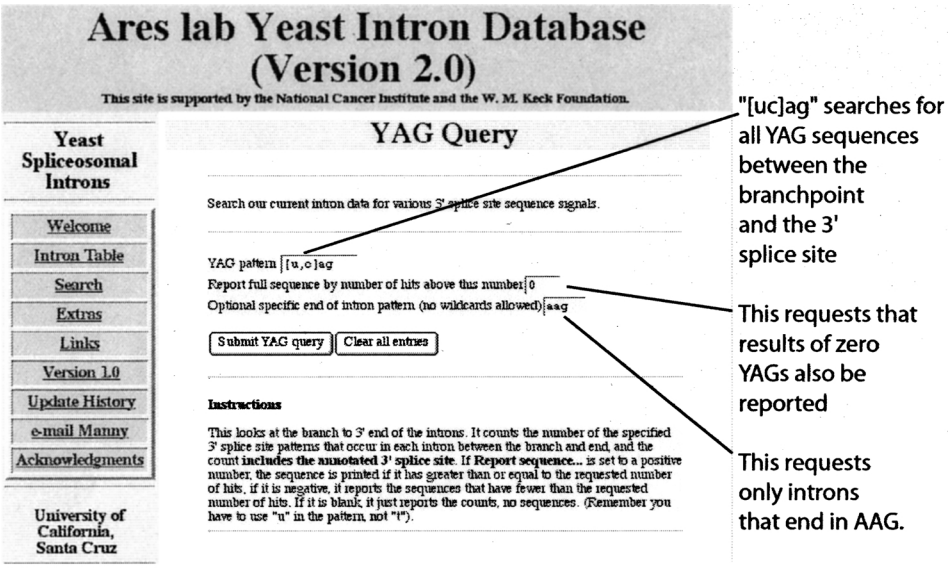


FIG. 5. Splice site query.

branchpoint and the true 3′ splice of each intron. Go back again to the "Search" page and click on the "YAG Query" link. This is the YAG Query form page and it is shown in Fig. 6. In the first box, type the nucleotide sequence pattern to search for between the branchpoint and the 3′ splice site. In the example shown, "[uc]ag" (use square brackets, not braces or parentheses!), the computer will search for all pyrimidine-A-G sequences found between the branchpoint and the 3′ splice site. If nothing is put into the "Report full sequences . . ." box, then only the counts of the number of introns in each class will be returned. To see the full report (bottom of Fig. 6), put "0" (the number zero) in that box. If you are only interested in introns that have one or more occurrence of the sequence, put "1." In our search here, we have also restricted the introns we want to look at to be the ones that end in the unusual AAG 3′ splice site. This is specified by including the last three bases in the desired introns in the "Optional specific end of intron pattern" box. One can specify more or fewer bases using this box as well (e.g., "UUAAG"), but it does not accept wild-card characters or bracket requests.

The output of the search is shown in the bottom half of Fig. 6. The first line contains the number of occurrences of the pattern, the next line has links to the report for



The output looks like this:

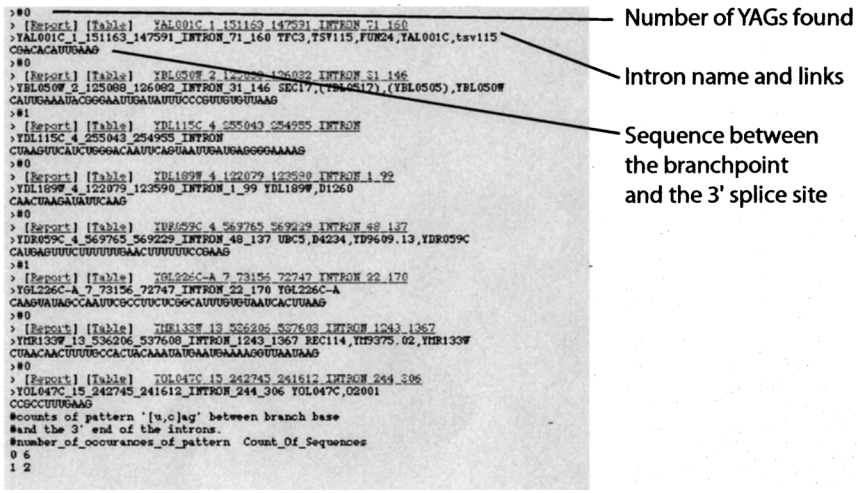


FIG. 6. Using the YAG query.

that intron, including the gene name(s), and the last line has the sequence between the branchpoint and the 3′ splice site. The search shows that two genes, *YDL115C* and *YGL226C-A*, have introns that skip a YAG sequence in favor of an AAG sequence. In the case of *YGL226C-A*, the UAG is only 9 nt from the branchpoint, a distance less than that found in the intron with the shortest such distance (*MATa1*

second intron, 10 nt). Care must be used in evaluating the sequence by eye, since in about 16 introns there is a CAG immediately after the branched residue, and these are unlikely to represent 3′ splice sites. (Can you find these? *Hint:* Go back to the Intron Splice Signals Query page and type "cag" into the "Bpost" box and submit the form.) The search for sequences on the YAG Query page is not limited to 3′ splice site-like sequences. Actually any sequence can be requested using the "YAG pattern" box on the form. We have also used this search page to evaluate the existence of pyrimidine tracts and other sequences in the branchpoint-3′ splice site interval of yeast introns.

The final search is the Intron Sequence Search. Go back once more to the Search page and click on the "Sequence search on introns" link. This page (Fig. 7)

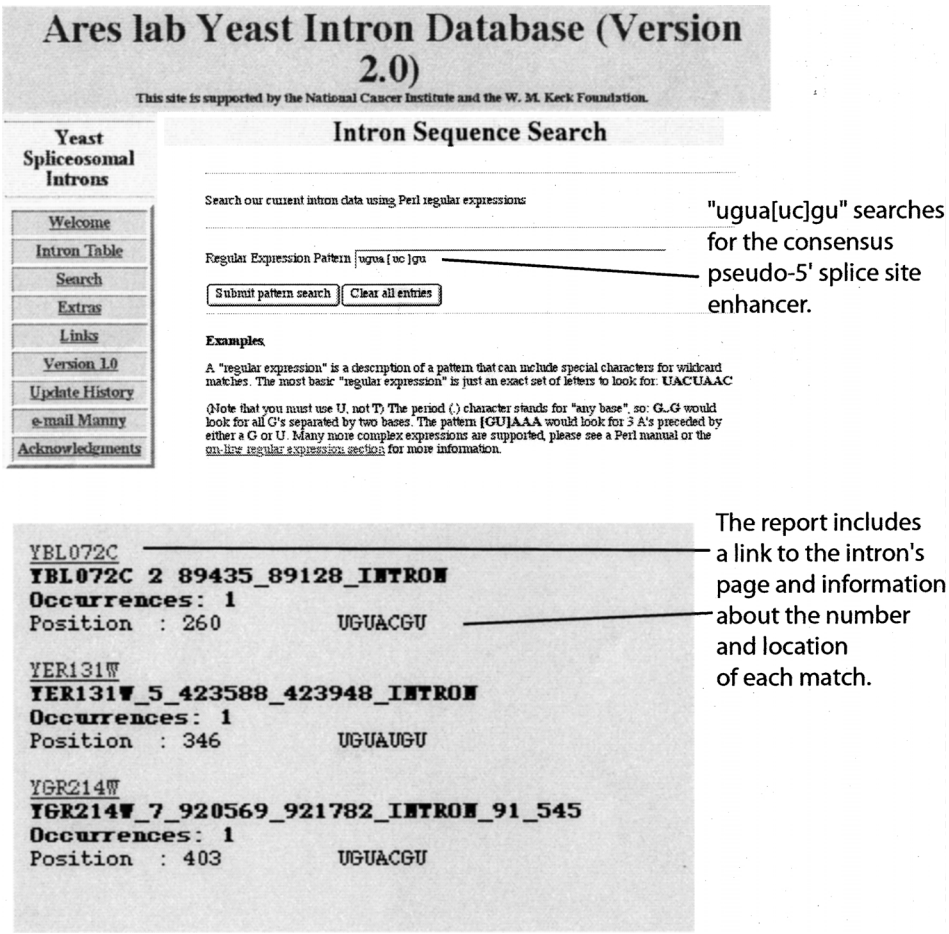


FIG. 7. Intron sequence search.

allows searches for any sequence anywhere in the intron. Type the sequence you are interested in finding into the box labeled "Regular Expression Pattern." You must use U instead of T! (Is our RNA bias showing?) This form accepts wild-card characters and bracket requests (square only!) within a string of nucleotides, such as "ugua[uc]gu." This particular request searches for either of two sequences: UGUAUGU or UGUACGU. Inserting a period "." means "any base" and is formally the same as "[agcu]." To get any pyrimidine at a particular position, as above, type "[uc]"; to get any purine, type "[ag]"; or for only G or C, use "[gc]." The "caret" character "^" can be used to specify "NOT" inside the brackets, so that if a nucleotide is excluded from a position this would be indicated by [^c], meaning "any nucleotide except C," or "A or U or G." To exclude two nucleotides, type "[^ag]." Note that "[^ag]" means the same to the computer as "[cu]" (you have just learned a little of the computer language Perl).

In the example in Fig. 7, we searched for the consensus sequence found in an *in vitro* evolution experiment that identified a pseudo-5′ splice site as an enhancer of splicing efficiency in yeast.[8] We wanted to see if such sequences might be common in yeast introns. To begin the search, we used the strictest definition of the consensus, which is essentially the same as the 5′ splice site consensus sequence except that it has a U upstream of the first G: UGUAYGU, which we express for the search as "ugua[uc]gu." Clicking on the "Submit pattern search" button, the computer returns a page containing information on three introns. Note that the natural 5′ splice site is excluded because the intron sequence data in the database includes no exon nucleotides, and thus none has a U upstream, at least in the data being searched. Many more introns are returned if the first U is deleted from the request. The results provide a link to the report page, the intron coordinates, the number of occurrences of the sequence, and the position of the sequence found. These three introns might be good candidates in which to test the idea that pseudo-5′ splice site sequences contribute to splicing in natural yeast introns. Additional searches with more relaxed consensus sequences may reveal additional candidates.

Database Errors, Programming Bugs, and Interpretational Caveats

The Ares lab Intron Database is a work in progress. We originally devised it for our own use, but found it so useful that we thought others might want to use it as well. We try to maintain the accuracy of the data, but there is a lot of it underneath. (One can think of the distinction between a searchable database and a publication as similar to the distinction between performance art and painting. Each time you access the database, we are putting on a show, which is a different responsibility than one has after painting a picture.) Although the data are useful for gaining broad-brush impressions of the intron family in yeast, and the choice nugget will occasionally be found, all specific results of importance should be

Au: lc ok l?

[8] D. Libri, A. Lescure, and M. Rosbash, *RNA* **6,** 352 (2000).

confirmed by comparing the findings with information in other databases. SGD, PubMed, and GenBank employ individuals who are responsible for constantly updating information of relevance to the data in our collection. Database errors exist in all databases, so beware! Also note that future versions of the database may return results slightly different from those presented in the figures, because of updated information.

There are a few programming bugs yet in the system as well. The Intron Sequence Search report gives spurious position information if there is more than one occurrence of a particular sequence (although the first listed position is correct, and the other occurrences are present) and we are working to fix that. One limitation we cannot really surmount involves the identification of the branchpoint. We specify this position based on looking at the sequence and in many cases it is a guess. Molecular analysis of branchpoints is challenging, and there are few hard data for most introns. Finally one must be careful in interpretation of the findings. An example is the abundance of U tracts in introns. These are easy to find, and some introns have many, many short runs of U. The results of such a search may be impressive but the software assesses no significance to these results. The investigator must ask, How big is this intron? Given the G+C content of the yeast genome, how often might I expect to observe U runs of this length in a sequence of this size? What is the probability that my observation could be due to chance? In the future when all possible experiments have been done and all data is archived in searchable structures, all we will need do to test a hypothesis is to submit a form. Until then we can at least use our current databases to sharpen a few of our experimental rationales and hone some of our conclusions.

Acknowledgments

We thank Chuck Sugnet, Tyson Clark, Carrie Davis, and Marc Spingola for help keeping the database tidy. We also thank Haller Igel for comments on the manuscript. This work was supported by the W. M. Keck Foundation grant to the RNA Center at Santa Cruz, and by grants to M.A. from the National Institutes of Health (CA 77813 and GM 40478).