888–894 Nucleic Acids Research, 1999, Vol. 27, No. 3

Genomic detection of new yeast pre-mRNA 3′-end-processing signals

Joel H. Graber*, Charles R. Cantor, Scott C. Mohr[1] and Temple F. Smith[2]

Center for Advanced Biotechnology, Boston University, 36 Cummington Street, Boston, MA 02215, USA, [1]Department of Chemistry, Boston University, Boston, MA 02215, USA and [2]Biomolecular Engineering Research Center, 36 Cummington Street, Boston, MA 02215, USA

Received August 24, 1998; Revised and Accepted December 13, 1998

ABSTRACT

To investigate *Saccharomyces cerevisiae* 3′-end-processing signals, a set of 1352 unique pre-mRNA 3′-end-processing sites, corresponding to 861 different genes, was identified by alignment of expressed sequence tag sequences with the complete yeast genome. Nucleotide word frequencies in the vicinity of the cleavage sites were analyzed to reveal the signal element features. In addition to previously recognized processing signals, two previously uncharacterized components of the 3′-end-processing signal sequence were discovered, specifically a predominance of U-rich sequences located on either side of the cleavage site. One of these, the downstream U-rich signal, provides a further link between the 3′-end-processing mechanisms of yeast and higher eukaryotes. Analysis of the complete set of 3′-end-processing sites by means of a discrimination function supports a 'contextual' model in which the sum total effectiveness of the signals in all four elements determines whether or not processing occurs.

INTRODUCTION

The complete genome sequences now available afford new avenues to analyze experimental data, for example, the databases of expressed sequence tags (ESTs). In addition to genetic expression levels as functions of cellular conditions (1,2), these databases contain potential information about pre-mRNA processing mechanisms, particularly signal sequences, and choices among alternate splicing or alternate end-processing sites (3). While regulatory mechanisms operative at the 5′ ends of RNA transcripts have generated much interest, these 3′-end features have been less thoroughly investigated. Sequencing of cDNA generated with an oligothymidylate primer, which hybridizes to the polyadenylate tail found at the 3′-end of mature mRNA yields 3′-ESTs. Comparison of 3′-EST and genomic sequences defines the 3′-end-processing (cleavage and polyadenylation) site as the location of the 3′-most end of common sequence. Through statistical analysis of the genomic sequences surrounding such processing sites, we can identify signals used to activate the 3′-end-processing mechanism. This informatic approach complements and extends the traditional biochemical methods by examining a much larger number of genes.

The *Saccharomyces cerevisiae* 3′-end-processing signal has been reviewed previously (4–7). It was proposed to consist of three elements: the cleavage site, a 'positioning' element, located 10–30 nucleotides (nt) upstream of the cleavage site, and an 'upstream efficiency' element, usually located another 10–30 nt upstream of the positioning element, though this separation was highly variable.

MATERIALS AND METHODS

We obtained a collection of 3425 *S.cerevisiae* EST sequences from the public ftp server at The Institute for Genome Research (TIGR), and located these sequences within the complete yeast genome (8), as described elsewhere (J.H. Graber, C.R. Cantor, J.M. Freeman, T.N. Plasterer and T.F. Smith, manuscript in preparation). The 3266 ESTs that could be uniquely placed within the genome were compared with the annotated open reading frames (ORFs) (9) for possible identification as 5′- or 3′-ESTs of a particular gene. Chromosome sequences and ORF descriptions were obtained from the Saccharomyces Genome Database (10).

To determine the signals for 3′-end processing, 1352 unique, unambiguous 3′-EST sequences were extracted from the total. These ESTs were required to be unique, to avoid skewing of statistical data, and unambiguously located at the 3′-most end of the mRNA. EST sequences were eliminated from signal analysis for the following reasons: (i) lack of, or ambiguous, EST–ORF association (231 ESTs eliminated); (ii) identification as a 5′-EST (475 ESTs); (iii) occurrence in the genomic sequence of the recognition sequences for either of the restriction endonucleases (*XhoI* and *Eco*RI) that were used to insert the ESTs into the cloning vector (121 ESTs); (iv) identification as a 3′-EST with cleavage position preceding a stop codon (40 ESTs); (v) occurrence of an A-rich region (at least six A residues in the next 10 positions) immediately on the 3′ side of the putative cleavage site. Elimination of 3′ ESTs with A-rich regions possibly excludes valid 3′-end-cleavage sites; however, the determination is ambiguous since such an A-rich region is a potential

* To whom correspondence should be addressed. Tel: +1 617 353 8500; Fax: +1 617 353 8501; Email: jhg@darwin.bu.edu

hybridization site for the oligo-(dT) primer used in EST generation (620 ESTs); or (vi) duplication of putative cleavage positions (as indicated by the 3′-end of the EST-genome match) (427 ESTs).

For characterization of signal elements, we chose to work with six-letter words since, given the size of our data set, over-representation (indicating a signal) among words of greater length would have only weak statistical significance. In addition, most previously reported signal elements (4) are ≤ 6 nt in length. The predicted abundance of 6-nt words was computed using the measured dinucleotide frequencies for all 1352 end-processing sequences from positions −100 to +50 with respect to the putative cleavage position. For example

$$q(\mathrm{TATATA}) = p(\mathrm{T}) * p(\mathrm{TA|A}) * p(\mathrm{AT|T}) * p(\mathrm{TA|T}) * p(\mathrm{AT|A}) * p(\mathrm{TA|T}) \quad \mathbf{1}$$

where $p(\mathrm{T})$ is the measured frequency of T in any position, $p(\mathrm{TA|T})$ is the measured frequency of A following T, and $p(\mathrm{AT|A})$ is the measured frequency of T following A.

We determined the statistical significance of potential signal-element words by comparing their measured, non-self-intersecting, abundance (p) with the predicted abundance (q), as defined above. We ranked the 6-nt words based on log likelihood $[p * \ln(p/q)]$, which has the advantage of being robust against random occurrences of extremely improbable words.

In order to determine the optimal signal elements, we used an iterative filtering technique, in which we found the statistically most significant words in each signal-element region (Fig. 4B), ranked by log-likelihood. In each successive iteration, the sequences were clustered on the basis of presence of the optimal (highest ranked) words in each signal element, arbitrarily defined as words with log-likelihood values ≥80% of the highest value. For each signal element, the sequences that did not contain any of the optimal words in other three signal-element regions were used to generate a new log-likelihood ranking of the signal words. This procedure was iterated to approximate convergence upon a constant set of words.

We developed a quantitative discrimination function to identify potential 3′-end-processing sites. Candidate sequences are partitioned into signal-element regions (Fig. 4B), and then each region is searched for the words identified for that element by the iterative filter. The discrimination function is a linear combination of the highest log-likelihood found for each of the signal elements and the local GC-content, independent of the four signal words. (The GC-content was included in the score based on the reduced GC-content apparent in the 3′-UTR regions of Fig. 2A and B.) 1475 random sequences taken from known coding sequences were used as a negative training set. Covariance analysis (11) was used to optimize the coefficients for discrimination between the EST-indicated cleavage sites and the coding sequences. The sets of coefficients give each candidate sequence five values that can be used as coordinates in a five-dimensional space, in which probable 3′-end processing sites form clusters largely distinct from those of the negative training set (Fig. 5A).

Throughout this paper, we use the convention of naming nucleotides based on DNA residues (A, C, G, T) rather than RNA residues (A, C, G, U) since all of our analysis was performed using *S.cerevisiae* genomic DNA sequence.

RESULTS

We identified 1352 unique unambiguous 3′-end ESTs as described in Materials and Methods. Figure 1 shows the distribution of 3′-untranslated-region (3′-UTR) lengths associated with the processing sites. The 1352 processing sites represent 861 unique genes; the maximum number of cleavage sites for a single gene is 10, and the largest range over which cleavage occurs in a single gene is 454 nt.

We aligned the 1352 unique sequences on the putative cleavage sites, and measured the single-base frequencies for all positions within 100 nt on either side. The resulting plot is shown in Figure 2A, where position 0 is the putative cleavage site. Non-random distributions of all 4 nucleotides flank the spike corresponding to the A residue at the cleavage site. As a control, we repeated this operation for each of the 861 unique genes with the sequences aligned on their stop codons rather than 3′-cleavage sites, as shown in Figure 2B. The stop codon (TGA, TAA, TAG) stands out clearly at position 0, as does the 3-nt oscillation of the preceding coding sequence. Randomly aligned 3′-UTR sequences (data not shown) produce a distribution similar to that shown at the extreme right (position greater than 40) of Figure 2A. We used single-base frequencies from the randomly aligned sequences as a background to measure the information content or statistical importance of each position of the aligned sequences. Figure 3 is a combination of the single-base frequencies and the Kullback–Leibler asymmetric divergence measure (12,13) for each position in the aligned sequences. Significant divergence from the background distribution is apparent from relative positions −75 to +25 with respect to the cleavage site.

In order to identify signal sequences, we measured the abundance of all 6-nt words in the sequences flanking the putative 3′-end-processing site. We first searched for the words with the highest log-likelihood scores (Materials and Methods), without regard to specific location within the sequence. Table 1 lists the 25 highest-scoring words across the entire 3′-end-processing region.

To further characterize the most significant words, we measured their abundance as a function of sequence position relative to the putative 3′-end-processing site. The most significant words can be grouped based on similarity of these positional abundance plots. Figure 4A shows three typical distributions of significant words. For clarity, only a representative sampling of

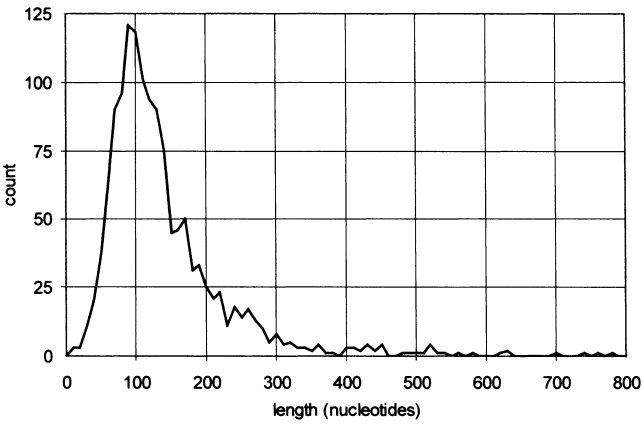

Figure 1. Distribution of 3′UTR lengths determined for 1352 unique 3′-ESTs from *S.cerevisiae*, selected as described in the text. Average length, 144 nt; median length, 121 nt.

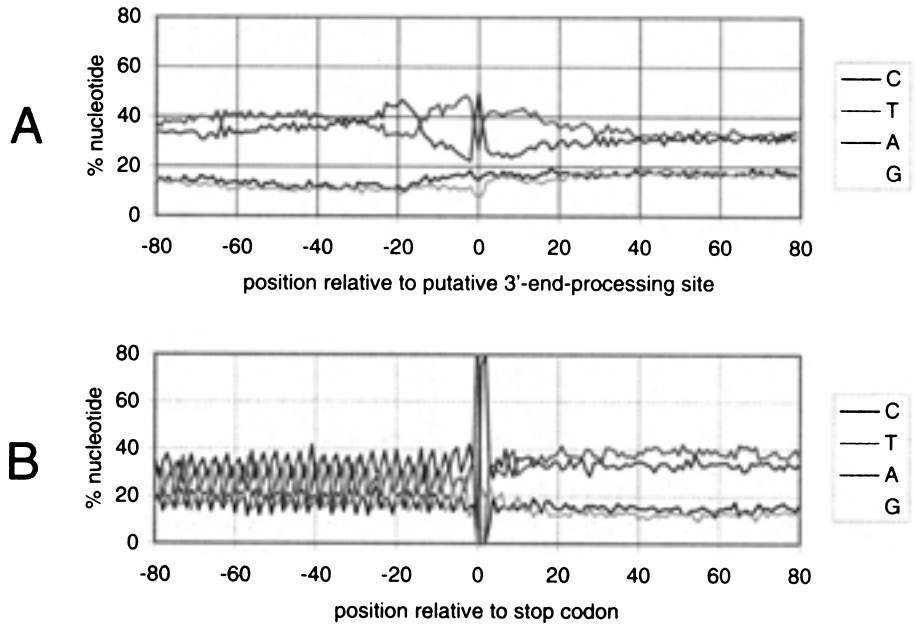

Figure 2. Single-nucleotide frequencies in 1352 unique *S.cerevisiae* 3′-end-processing regions aligned (A) on the putative cleaveage site and (B) on the stop codon. The average single-nucleotide frequencies for all non-coding sequences in the *S.cerevisiae* genome are C, 17.6%; T, 32.4%; A, 32.5%; G, 17.5%.

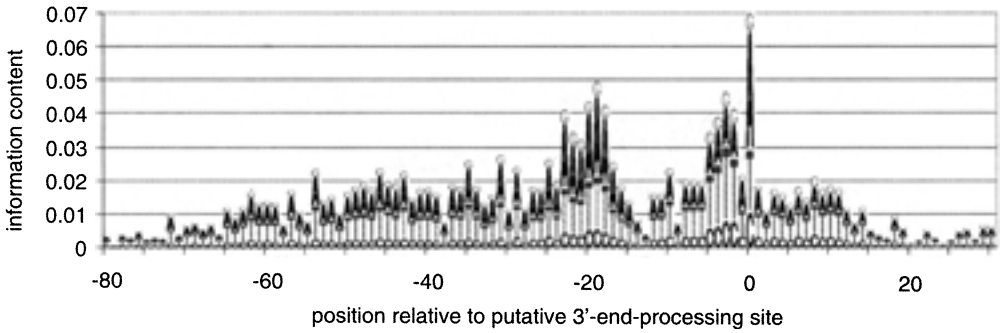

Figure 3. Statistical significance of each position in 1352 unique, aligned *S.cerevisiae* 3′-end-processing regions, plotted as Kullback–Leibler asymmetric divergence measure. This plot is similar to a Sequence Logo (23) plot, but does not assume a uniform background at each nucleotide. At each position, the total column height is a measure of divergence from background, while the heights of the individual nucleotide letters are proportional to their respective abundance.

each group is shown. Based on the measured distributions, we were able to define four signal-element regions, with boundaries as shown in Figure 4B.

The optimal signal elements, as determined by the iterative filtering process (described in Materials and Methods), are listed in Table 2. The filtering process contains the implicit assumption that the optimal signal elements will not commonly appear together. The assumption appears justified, since the optimal words for both the upstream efficiency and positioning elements determined by this method agree with the experimentally deduced signals for yeast 3′-end processing (14,15). From these results, we draw the conclusion that the presence of an optimal word in one signal element makes possible the use of suboptimal words in the other elements.

We used a discrimination function (described in Materials and Methods) as a means of differentiating between 3′-end-processing sites and random coding sequence, which we have assumed to be devoid of 3′-end-processing sites. Figure 5A and B shows two- and one-dimensional projections, respectively, of the five-dimensional discrimination score vectors determined for the 1352 3′-end-processing (positive) sequences, as well as 1475 random coding (negative) sequences. The horizontal coordinate in both plots is the dimension of maximum separation of the positive and negative sets.

DISCUSSION

Mutagenesis experiments (14) previously demonstrated that TATATA encodes the optimal sequence for the efficiency element. Our analysis confirms both the composition and position of this element, as shown in Tables 1 and 2 and in Figure 4A. In addition, while TATATA encodes the most commonly used signal, nearly all single-base-change transitional substitution mutations of this sequence (e:g. TACATA, TATGTA) are statistically significant and display the same positional distribution as the TATATA sequence. Mutagenesis experiments have also determined the optimal positioning-element sequence in yeast to be the higher eukaryote consensus sequence AATAAA, though this element tolerates wide variability (15). The consensus has

Table 1. Most significantly over-represented 6-nt words in 1352 *S.cerevisiae* mRNA 3′-end-processing regions

Word	Sequences with at least one occurence	Frequencies measured (p)	Predicted (q)	Log likelihood $p * \ln(p/q)$
TATATA	756	0.358	0.133	0.353
ATATAT	682	0.328	0.136	0.288
TTTTCT	593	0.257	0.107	0.224
TTTCTT	578	0.244	0.107	0.201
TTTTTC	531	0.235	0.113	0.173
CTTTTT	542	0.238	0.116	0.171
TGTATA	430	0.172	0.064	0.170
TTCTTT	525	0.223	0.107	0.163
TCTTTT	509	0.219	0.107	0.156
TATGTA	406	0.161	0.064	0.149
TACATA	377	0.158	0.062	0.148
AAGAAA	404	0.158	0.064	0.143
TTTTAT	665	0.317	0.212	0.127
ATTTTT	697	0.339	0.233	0.127
CATATA	359	0.148	0.063	0.127
TATTTT	626	0.312	0.212	0.121
ATATAC	360	0.139	0.059	0.119
ATATAA	547	0.222	0.134	0.113
GTATAT	340	0.135	0.060	0.110
ACATAT	345	0.137	0.063	0.106
GAAGAA	224	0.089	0.027	0.105
AAATAA	534	0.231	0.147	0.104
GAAAAA	349	0.142	0.069	0.103
AGAAAA	336	0.134	0.064	0.099
TTTTA	675	0.310	0.228	0.096

been described simply as 'A-rich' (4). Our analysis confirms this assessment in both content and position, as shown in Tables 1 and 2 and Figure 4.

The cleavage site has previously been described as a pyrimidine, followed by three or more adenines (4). Our results indicate that this is a limited model. The cleavage and polyadenylation preferentially occur prior to an adenine residue (Figs 2A and 3); however, we have also found a predominance of T-rich elements, positioned both immediately before and immediately after the cleavage site, that have not been previously associated with 3′-end-processing signals in yeast. The occurrence of T-rich elements is intriguing, since T-rich signal sequences have been implicated as encoding a U-rich downstream 3′-end-processing element in higher eukaryotes (5–7). Recent studies (16–19) of the protein components of the complexes involved in 3′-end processing indicate greater similarity between yeast and higher eukaryotes than had been previously recognized. The occurrence of T-rich sequences near or beyond the cleavage site in *S.cerevisiae* mRNA further underscores this similarity.

We believe that a model of 3′-end processing that takes into account both kinetic and thermodynamic effects of multiple and cooperative protein–RNA binding is necessary to explain fully the observed variations in signal-element combinations. Such a model would call for a 3′-end-processing signal determined by the sum total characteristics of all four signal elements for any specific pre-mRNA. A similar 'contextual' model was recently proposed to explain promoter activity in the absence of the consensus 'TATA'-box signals (20); The need for an imperfect version of a specific signal to fit into a local context of multiple, potentially overlapping signals has also been previously discussed (21).

Our model implies that suboptimal signals in one or more elements can be compensated for by strong signals in some or all of the remaining elements. Additionally, 3′-end processing can occur in the absence of any of the optimal signal words, as

Table 2. Top ranked 6-nt words for each signal element in yeast mRNA 3′-end-processing sequences, as determined by the iterative filtering procedure

Efficiency (I)		Positioning (II)		Pre-clevage (III)		Downstream (IV)	
Word	Score	Word	Score	Word	Score	Word	Score
TATATA	1.55	**AAAATA**	0.97	**TTTTAT**	0.72	**TTTTCT**	0.46
ATATAT	1.15	**AATAAA**	0.92	**TTTTTT**	0.72	**CTTTTT**	0.44
TATGTA	0.63	**ATAATA**	0.87	**TATTCT**	0.67	**TTTTTC**	0.44
TGTATA	0.62	TAATAA	0.77	**TTTCTT**	0.60	**TTTCAT**	0.37
TACATA	0.54	AATATA	0.77	**TTCTTT**	0.60	TATTCT	0.30
GTATAT	0.47	AAATAA	0.67	ATTTTT	0.55	TTCATT	0.30
CATATA	0.46	AAAAAA	0.62	TTTTTA	0.46	TTTATT	0.26
ACATAT	0.38	AAGAAA	0.59	TATTAT	0.46	TATTTC	0.25
ATGTAT	0.37	AAAAAT	0.57	TTCTTC	0.44	TCTTTT	0.24
ATATAA	0.37	ATAAAA	0.51	TTTTTC	0.42	TCATTT	0.24

The iterative filtering procedure is defined in the main text. The 'optimal' words are displayed in bold face. The Roman numeral signal element identifiers correspond to Figure 4.

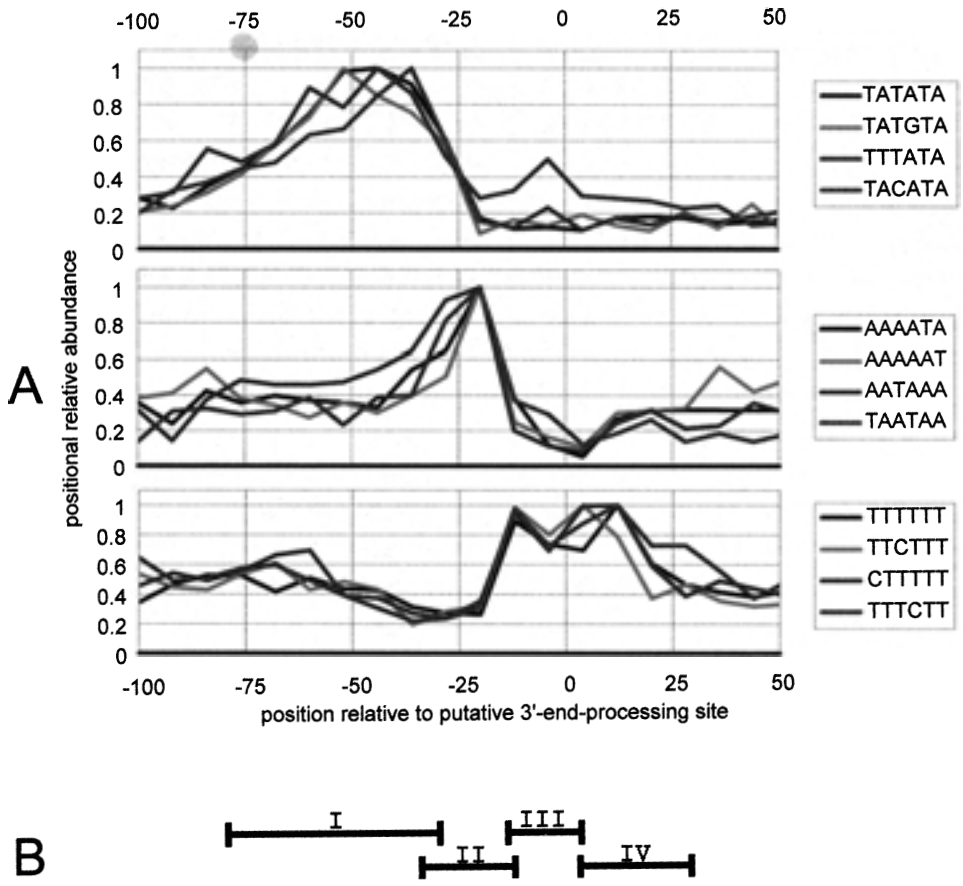

Figure 4. Positional distribution of statistically significant 6-nt words in 1352 unique, aligned *S.cerevisiae* 3′-end-processing regions. (A) Six-nucleotide word usage as a function of sequence position (relative to the putative cleavage site). Signal words were clustered based on similarity of the positional distributions, and a representative sample of each cluster is displayed. (B) Four signal-element position regions derived from the observed clusters shown in (A). The signal-element position regions were used as boundaries for the determination of the optimal signal-element words, as shown in Table 2.

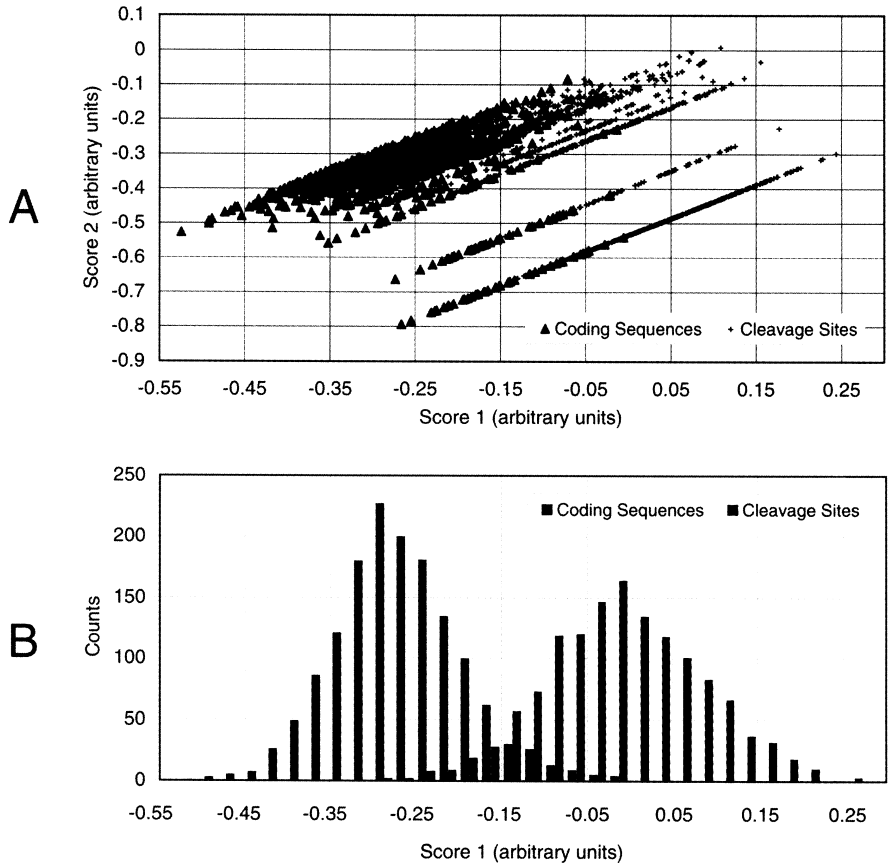

Figure 5. (A) Two- and (B) one-dimensional projections of the five-dimensional scoring space used to discriminate between *S.cerevisiae* 3′-end-processing sites and random coding sequences. The horizontal axis in both plots is the dimension of greatest discrimination, as determined by covariance analysis.

long as the total signal is adequate. Indeed, the iterative filtering indicates that >30% of the genes examined in this study are cleaved in positional contexts that use none of the optimal signal elements (defined in Table 2). The results of the iterative filtering also indicate the predominance of the efficiency element as the key component of the overall signal. Approximately 40% of the genes investigated use the optimal word (TATATA). The great majority of these have suboptimal signals in all other positions.

The two-dimensional projection (Fig. 5A) of the five-dimensional scoring function space appears to support a multiclass nature of the 3′-end-processing signals. Inspection of the discrimination function coefficients (data not shown) makes clear that the lower right hand group of sequences in Figure 5A corresponds to sequences that have TATATA as the efficiency element. The TATATA efficiency element is clearly a strong indicator of a true 3′-end-processing site; however, the significant separation between positive and negative sequences without the TATATA element supports our contention that 3′-end processing is determined by the net characteristics of the complete signal.

Our model also provides a potential explanation for the lack of previous detection of the T-rich signals near the cleavage site: sequences with optimal efficiency and positioning elements [TATATA and AATAAA (22), respectively] do not require strong signals in either the pre-cleavage or downstream elements. Close inspection of previous experimental studies reveals that, in many cases, T-rich elements were present, but not noted as significant (e.g. 14,15,22).

Our analysis exemplifies the insights to be gained through the use of whole-genome sequences in conjunction with experimental data. The integration of EST and genomic data provided two principal benefits: (i) we were able to eliminate over half of the EST sequences from our data set, since they were not unambiguous 3′-end-processing sites; and (ii) we gained additional sequence data not contained in the original ESTs. Without this additional data, the downstream T-rich element would have gone undetected.

Bioinformatic analysis, in contrast to exclusively experimental approaches, allows a much broader sampling of genes for signal sequences, thereby providing statistical data beyond the range of feasible experiments. The prior work, while limited in genomic scope, was a thorough exploration of signals that could function as 3′-end-processing signals. Our work, by contrast, is a study of the signals that are present across many 3′-end-processing regions. It is significant that, in the case of the efficiency and positioning elements, the two methods produced essentially the same results.

ACKNOWLEDGEMENTS

The authors thank K. Weinstock and J. C. Venter of The Institute for Genome Research (TIGR) for making the 3425 *S.cerevisiae* EST sequences publicly available. Tom Gilmore, Dean Tolan, Chip Celenza and Geof Cooper provided critical readings of the

manuscript. Jadwiga Bienkowska assisted with the covariance analysis. J.H.G. is supported by a training grant from the US National Human Genome Research Institute (T32 HG00041-03) and by Sequenom, Inc. S.C.M. is also partially supported by training grant T32 HG0041-03. T.F.S. is supported by grants from the National Library of Medicine (LM05205) and the Department of Energy (DE-FG02-98ER62558).

REFERENCES

1 Lashkari,D.A., DeRisi,J.L., McCusker,J.H., Namath,A.F., Gentile,C., Hwang,S.Y., Brown,P.O. and Davis,R.W. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 13057–13062.
2 DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) *Science*, **278**, 680–686.
3 Gautheret,D., Poirot,O., Lopez,F., Audic,S. and Claverie,J.M. (1998) *Genome Res.*, **8**, 524–530.
4 Guo,Z. and Sherman,F. (1996) *Trends Biochem. Sci.* **21**, 477–481.
5 Keller,W. and Minvielle-Sebastia,L. (1997) *Curr. Opin. Cell. Biol.*, **9**, 329–336.
6 Colgan,D.F. and Manley,J.L. (1997) *Genes Dev*., **11**, 2755–2755.
7 Manley,J.L. and Takagaki,Y. (1996), *Science*, **274**, 1481–1482.
8 Goffeau,A., Aert,R., Agostini-Carbone,M.L., Ahmed,A., Aigle,M., Alberghina,L., Albermann,K., Albers,M., Aldea,M., Alexandraki,D. *et al.* (1997) *Nature*, **387** (6632 Suppl.), 1–105.
9 Cherry,J.M., Ball,C., Chervitz,S., Dolinski,K., Dwight,S., Harris,M., Hester,E., Juvik,G., Malekian,A., Roe,T., Weng,S. and Botstein,D. (accessed March, 1, 1998) http://genome-www.stanford.edu/Saccharomyces/
10 Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldman,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and Oliver,S.G. (1996) *Science*, **274**, 546.
11 Rencher,A.C. (1995) *Methods of Multivariate Analysis*. John Wiley & Sons, Inc., New York.
12 Kullback,S. (1959) *Information Theory and Statistics*. Dover Publications, New York.
13 Kullback,S. and Leibler,R.A. (1986) *Ann. Math. Stat.*, **22**, 79.
14 Guo,Z. and Sherman,F. (1995) *Mol. Cell. Biol.*, **15**, 5983–5990.
15 Russo,P., Li,W.Z., Guo,Z. and Sherman,F. (1993) *Mol. Cell. Biol.*, **13**, 7836–7849.
16 Zhao,J., Kessler,M.M. and Moore,C.L. (1997) *J. Biol. Chem.*, **272**, 10831–10838.
17 Preker,P.J., Ohnacker,M., Minvielle-Sebastia,L. and Keller,W. (1997) *EMBO J.*, **16**, 4727–4737.
18 Stumpf,G. and Domdey,H. (1996) *Science*, **274**, 1517–1520.
19 Chanfreau,G., Noble,S.M. and Guthrie,C. (1997) *Science*, **274**, 1511–1514.
20 Audic,S. and Claverie,J.M. (1998) *Trends Genet.*, **14**, 10–11.
21 Trifonov,E.N. (1996) *Comput. Appl. Biosci.*, **12**, 423–429.
22 Guo,Z. and Sherman,F. (1995) *Mol. Cell. Biol.*, **16**, 2772–2776.
23 Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.