Nucleic Acids Research, 1993, Vol. 21, No. 16 3829–3838

A quality control algorithm for DNA sequencing projects

Owen White, Ted Dunning[1], Granger Sutton, Mark Adams, J.Craig Venter and Chris Fields*
The Institute for Genomic Research, 932 Clopper Road, Gaithersburg, MD 20878 and
[1]Computing Research Laboratory, Box 30001/3CRL, New Mexico State University, Las Cruces, NM 88003-0001, USA

Received March 29, 1993; Revised and Accepted June 24, 1993

ABSTRACT

Heterologous DNA sequences from rearrangements with the genomes of host cells, genomic fragments from hybrid cells, or impure tissue sources can threaten the purity of libraries that are derived from RNA or DNA. Hybridization methods can only detect contaminants from known or suspected heterologous sources, and whole library screening is technically very difficult. Detection of contaminating heterologous clones by sequence alignment is only possible when related sequences are present in a known database. We have developed a statistical test to identify heterologous sequences that is based on the differences in hexamer composition of DNA from different organisms. This test does not require that sequences similar to potential heterologous contaminants are present in the database, and can in principle detect contamination by previously unknown organisms. We have applied this test to the major public expressed sequence tag (EST) data sets to evaluate its utility as a quality control measure and a peer evaluation tool. There is detectable heterogeneity in most human and *C.elegans* EST data sets but it is not apparently associated with cross-species contamination. However, there is direct evidence for both yeast and bacterial sequence contamination in some public database sequences annotated as human. Results obtained with the hexamer test have been confirmed with similarity searches using sequences from the relevant data sets.

INTRODUCTION

Partial sequencing of randomly-selected cDNA clones to generate expressed sequence tags (ESTs) has become the method of choice for rapidly identifying new genes and characterizing transcript populations in tissues (1). EST projects are now underway for many human tissues (1–5) and for a variety of other organisms, including mouse (6), and *Caenorhabditis* (7, 8). A comprehensive EST sequence database (dbEST) has been established at the National Center for Biotechnology Information (USA) to maintain and distribute these data. The quality of the cDNA library from which clones are selected is critically important to the success of an EST project. Library quality has traditionally been assessed by clone diversity. A library that contains at least one full-length copy of a clone of interest is regarded as of good quality if the assay is screening for that particular clone. By sampling a library at random, EST projects can reveal many previously-unrecognized deficiencies (1, 2, 9, 10): clones without inserts, chimeras, unspliced messages, and contamination by clones from heterologous sources. When such contaminating clones can be identified by sequence similarity searches and discarded, no problem exists. The majority of clones sequenced in current EST projects, however, are not currently identifiable, even to gene family, by sequence similarity methods. Contaminating clones that are not identified as such, especially clones of heterologous origin, can lead to substantial errors in data interpretation.

As genome projects and biodiversity projects examine sequences from additional organisms, tissues, and populations, the problem of identifying heterologous clones will become progressively more serious. Medical tissue samples, for example, may contain previously uncharacterized protozoan parasites or other organisms which would contribute nucleic acids to any library prepared using conventional methods. Many plant and insect tissues are similarly populated by endophytes or parasites, many of which many never have been described, let alone sampled by sequencing. The problem of identifying heterologous clones in libraries is not limited to EST or other cDNA-based strategies. The use of large, not fully sequenced vectors, such as yeast or bacterial artificial chromosomes (YACs or BACs) as sources of DNA in genomic sequencing projects raises similar possibilities of contamination by unknown heterologous sequences. Such heterologous sequences cannot, in general, be identified by conventional experimental or computational means. Bacteria, protozoa, and fungi are all extremely diverse groups, with sequence differences within members of these groups often larger than those between plants and animals (11, 12). Hybridization of a cDNA library with total *E.coli* or *S.cerevisiae* DNA or RNA, for example, might identify contaminating sequences from these or closely related organisms, but would not detect all bacterial or fungal contaminants. The technical difficulties associated with such screening procedures are, moreover, very substantial and they are not routinely used. Sequence similarity searches of the public databases are very

* To whom correspondence should be addressed

straightforward, but are an even less sensitive means of detecting contaminants; only sequences that are significantly similar to already-sequenced genes can be detected. While the eventual complete sequencing of *E.coli*, *S.cerevisiae*, and other organisms will improve the odds of detecting heterologous sequences by similarity methods, sequences from distantly-related contaminants will still be difficult if not impossible to identify.

We have developed a rapid, quantitative statistical method that indicates the presence of heterologous sequences, even when they cannot be detected by sequence similarity searches, by capitalizing on the extensive differences in the hexamer composition of DNA from different organisms. Differences in the nucleotide, dinucleotide, and trinucleotide (codon) composition of DNA from different organisms are well known (13–16). With the availability of large data sets, statistically significant differences in the composition of oligomers, up to hexamers, have been observed between different organisms, and between coding exons and introns (15–17, O.White, T.Dunning and C.Fields, submitted). These differences are typically very pronounced between species of different kingdoms or phyla, and become progressively less significant for more closely related species. Comparison of the observed oligomer content of a sequence with the oligomer content of DNA from another organism can be used to determine the likelihood of that individual sequence being derived from that organism. When applied to all of the sequences from a genomic or cDNA library, such comparisons can provide an assessment of sequence heterogeneity in the library, and possibly of overall library quality. Here we describe an application of this new method to three *C.elegans* and six human EST data sets that have been either published or placed in the public sequence databases.

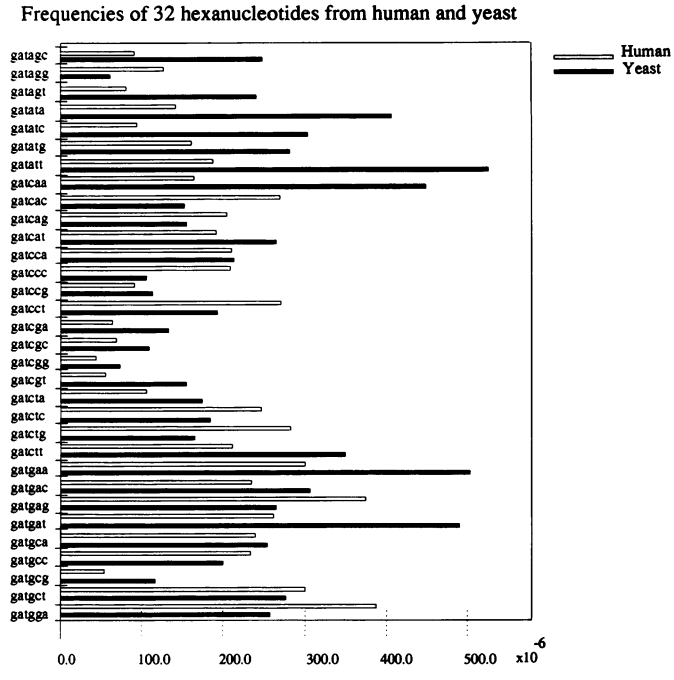

Figure 1. A representative sample of the variation in hexamer content between the sequences of two species. Shown is a histogram of raw frequencies of the 32 hexamers gatagc to gatgga from the large samples of exon and intron nucleotide sequences from *S.cerevisiae* and human used as controls in this work. As shown in figures 2–7, hexamer differences such as these can be used in the statistically significant separation of the sequences from one organism from those of another.

METHODS

The information content of an oligomer increases with oligomer length, until a length at which all longer oligomers are unique to the tested sequence is reached. Oligomers of length six (hexamers) provide a working compromise between the desirability of informative oligomers and the increased sampling uncertainty and impractically long execution times that result when oligomers closer to the uniqueness length are employed. The methods we have developed will work, however, with oligomers of any length, with sensitivity decreasing as length decreases, and sampling uncertainty and computational time increasing as length increases.

The simplest method of measuring hexamer composition is to count the number of occurrences of each of the 4096 possible hexamers in a sample sequence. Sequences from different organisms typically yield very different hexamer counts, as shown in Fig. 1. Hexamers, like individual nucleotides, can be viewed as being arranged randomly in DNA sequences; therefore, a statistical measure is needed to quantify the differences between hexamer counts from different sequences. A sequence of length L contains L-5 hexamers. The *a priori* probability of finding any particular hexamer in a sequence of less than 4 kilobases (kb) is [L-5]/4096. EST sequences are typically 300 nucleotides in length; any hexamer, therefore, has a 7.2% *a priori* probability of occurring in a typical EST. Our hexamer composition test uses a likelihood ratio measure that provides an accurate method for comparing a data set containing rare events to the contents of a large 'control' data set (18). For an EST sequence X and a large sample A of sequences from an organism of interest, the likelihood ratio $\lambda(A,X)$ is the ratio of the probability of finding a particular hexamer in X to the probability of finding that hexamer anywhere in the control set A. The log-likelihood ratio:

$$D(A,X) = -2\log\lambda(A,X) \quad (1)$$

is a measure of the dissimilarity of X to A; a large value of $D(A,X)$ indicates that the hexamer composition of X is very different from that of A. In practice, one question to ask is whether an EST is derived from its purported source organism, or from some heterologous source. If A is a control set of sequences from the purported source and B is a control set of sequences from a phylogenetically distant outgroup, the values:

$$Test(A,B,X) = D(A,X) - D(B,X) \quad (2)$$

will tend to be positive for test sequences that are less similar to the sequences from the putative source organism than to those from the heterologous outgroup. The cumulative distribution of $Test(A,B,X)$ will have an S-shaped signature, rising from an asymptotic value of zero to the left of the abscissa (where $Test(A,B,X) = 0.0$) to an asymptotic value of one to the right of the abscissa.

The values of $D(A,X)$ are calculated using a simple counting procedure. Let P_i refer to an array whose elements $P_{i...n}$ each contain a separate count for each hexamer encountered in the sequence. The elements of the array P_i correspond to the alphabetically arranged hexamers, so element $i=1$ contains the count for AAAAAA, element $i=2$ contains the count for AAAAAC, and element $i=4096$ contains the count for TTTTTT. Thus, the sequence AACCGGTTAACCGGTT has the following counts for each hexanucleotide:

element (i)	hexamer	count
91	AACCGG	2
364	ACCGGT	2
1456	CCGGTT	2
1725	CGGTTA	1
2801	GGTTAA	1
3010	GTTAAC	1
3095	TAACCG	1
3846	TTAACC	1

The distance measure $-2\log\lambda(A,X)$ (Eq. 1) is calculated using the following formula:

$$-2\log\lambda(A,X) = 2\times[\log L(A,A)-\log L(AX,A) + \log L(X,X)-\log L(AX,X)] \quad (3)$$

The expression in Eq. 3 denotes the sum of arrays A and X, i.e.: $AX_i = A_i+X_i$. The function $\log L(P,Q)$is a conventional log-likelihood ratio, where P and Q are any two sets of sequences, and:

$$\log L(P,Q) = \sum_i[(P_i\times\log Q)/\sum_i P_i] \quad (4)$$

The measure $-2\log L(A,X)$ is a distribution-independent measure that is approximated by Pearson's χ^2 statistic, but is accurate for rare events (18).

The hexamer counts for each test sequence are obtained and the values of *Test(A,B,X)* are calculated independently of the results for any other test sequence. The utility of this method as an assay for hexamer composition is, therefore, independent of the size of the test set. However, the confidence with which the results can be interpreted as indicating the presence of heterologous sequences in a library will depend on the size of the test set.

The cumulative distribution plots shown below were generated by sorting the values of *Test(A,B,X)* for each test set X. The N values for each test set X were plotted against a vertical axis obtained by assigning a vertical increment of $1/N$ to each successive point. Because the value of the*Test(A,B,X)* function depends on the length of the sequence, sequences that are greater than 400 nucleotides in length are trimmed to 300 nucleotides by a random window method before analysis. Results obtained with sequences shorter than 100 nucleotides or sequences with more than 2% ambiguous nucleotides, and hence ambiguous hexamers, may be difficult to interpret.

The DNA sequences used as controls for the EST data sets investigated here were extracted from GenBank. A single 95 kb sequence of *E.coli* genomic DNA (19), cut into nonoverlapping 300 nucleotide segments, was used as the bacterial control. The 315 kb sequence of *S.cerevisiae* chromosome III (20), cut into 300 nucleotide segments, was used as the yeast control. A total of 406 kb of genomic DNA sequences from 12 *C.elegans* cosmids (21, 22) was cut into 300 nucleotide segments for the *C.elegans* genomic DNA control. The human control set was a collection of sequences from GenBank that totalled 406 kb exclusively from the coding portions of exons and totalled 1,238 kb from introns (extracted as described in O. White, T. Dunning and C. Fields, submitted). Only exons and introns with no apparent annotation errors, or for which correct boundaries could be identified by examining the original publications were used. A total of 156 kb of coding exon sequences from *C.elegans* were used as the transcribed-sequence control in Fig. 4. Using randomly selected windows, the worm exons and the human control data that were used as test set sequences were trimmed to 300 nucleotides.

The analysis software is written in C. Comparison of a typical data set of a few thousand EST sequences with two control sets required one hour on a workstation (Sun SPARC 2). Source code is available by anonymous ftp from **ftp.tigr.org**; contact **owhite@tigr.org** for additional information.

RESULTS

Control data sets

The cumulative distributions of control sets of *E.coli* , *Saccharomyces cerevisiae*, *C.elegans*, and human sequences extracted from GenBank are shown in Fig. 2. The *E.coli* , *Saccharomyces*, and worm control sets are genomic DNA sequences; the human control set is a collection of sequences that are exclusively from the coding portions of exons or from introns. The cumulative distribution *Test(A,B,X)* for hexamer composition was calculated for each pair of control sets. Randomly chosen subsets of the sequences from one control set were removed from that control set and treated as the test set X. The *Test(A,B,X)* function is sensitive to sequence length. Because this is a statistical test requiring representative samples, the test sequences used to develop the control distributions each comprise roughly the same total number of hexamers. Therefore, all test set sequences were trimmed to a length of 300 basepairs (bp), the average size of most EST sequences. Each control distribution is the accumulated result of 20 repeated tests using random subsets of trimmed sequences. These distributions of hexamer composition tests show the upper and lower limits of variation in sequences from each control set. All of the distributions cross zero, indicating that each control set contains some sequences with hexamer compositions that are more similar to those of the outgroup than to other sequences from the source organism. Hexamers that are rare in one organism may be relatively common in a different organism; sequences containing even one rare hexamer may, therefore, score as quite dissimilar to the control composition of their source. Multiple independent calculations have been run using different randomly selected subsets of each control set; comparison of the results of these trials shows that the distributions shown in Fig. 2 are indeed representative of the control sets. Multiple trials in which the sequences in the selected subset X are all complemented before calculating *Test(A,B,X)* have also been run; the results of these complementation trials showed that the results of the hexamer composition test are independent of sequence orientation.

The utility of the test was investigated using test sets constructed by combining control sequences from human and *E.coli* , and from human and bacteria other than *E.coli* . As expected, these test sets yield distributions in which the human sequences overlap the human control and bacterial sequences overlap the bacterial control. Because each sequence in the test set is assayed independently, the extent of admixture of sequences from different sources has no effect on its detectability.

The distributions shown in Fig. 2 overlap for all pairs of control sets. The extent of overlap between two control distributions indicates the degree to which those sequence sets from those organisms are indistinguishable by measurement of hexamer composition. Control set sequences that have little overlap with a control set from some source organism are useful as outgroups in analyses of sequences from that organism. The overlap between

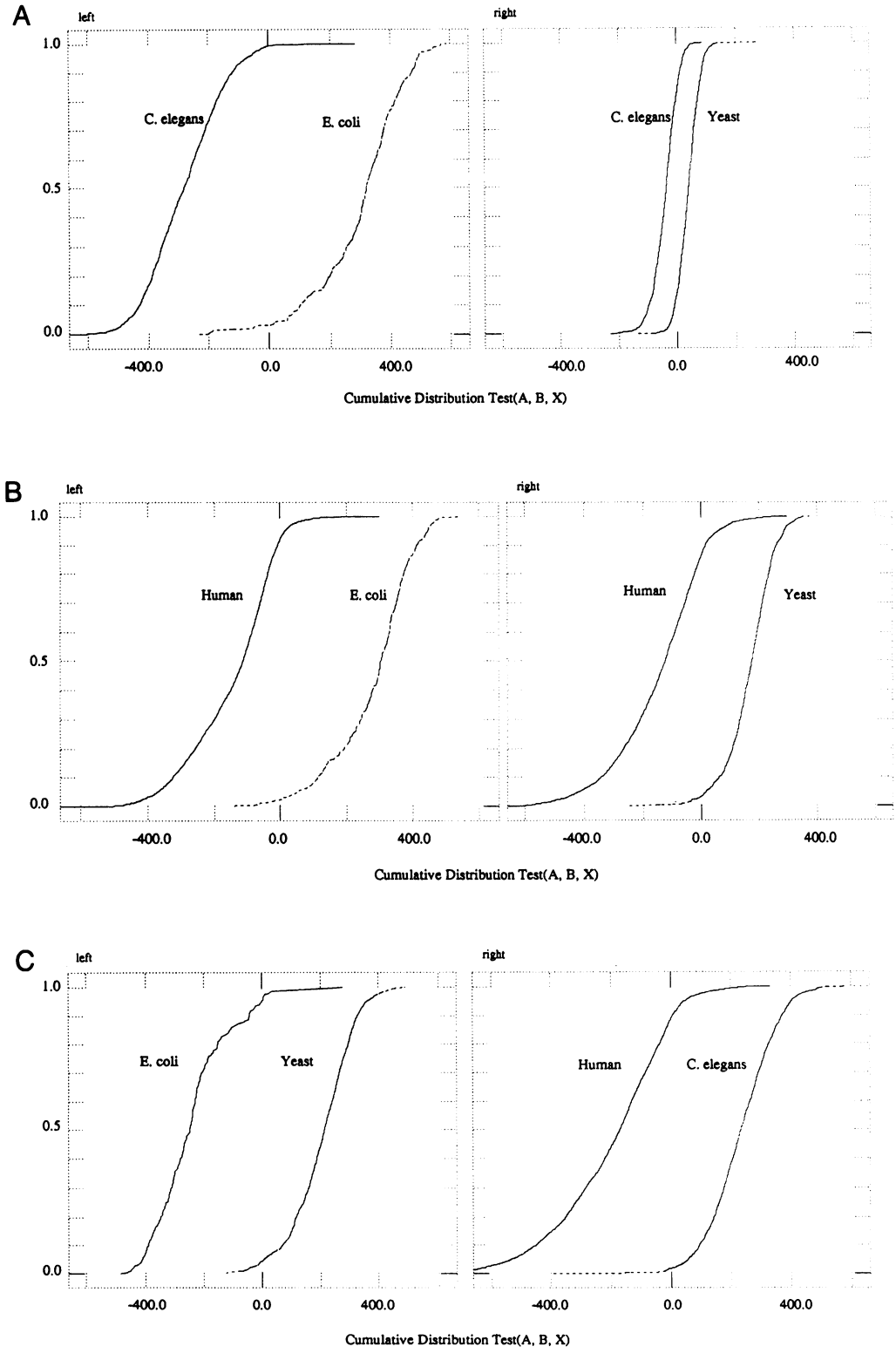

Figure 2. Separations between known human, *Caenorhabditis*, *Saccharomyces*, and *E.coli* sequences obtained from GenBank achieved using the test function for hexamer content. Plots of the normalized distributions of *Test(A,B,X)* were generated as described in Methods. Plots such as these allow all moments of the distribution to be compared visually, and hence, provide a more sensitive view of the data than average difference scores between the two distributions. The *Test(A,B,X)* values in an individual distribution are a measurement of dissimilarity of the sequences in the set *X* to those the two control data sets *A* and *B*. Each point of each distribution corresponds to a single test sequence. Provided that the control sets are representative of the source organisms, sequences falling outside of the area where the two control distributions overlap can be considered to have significantly different hexamer composition from that of the other control set.

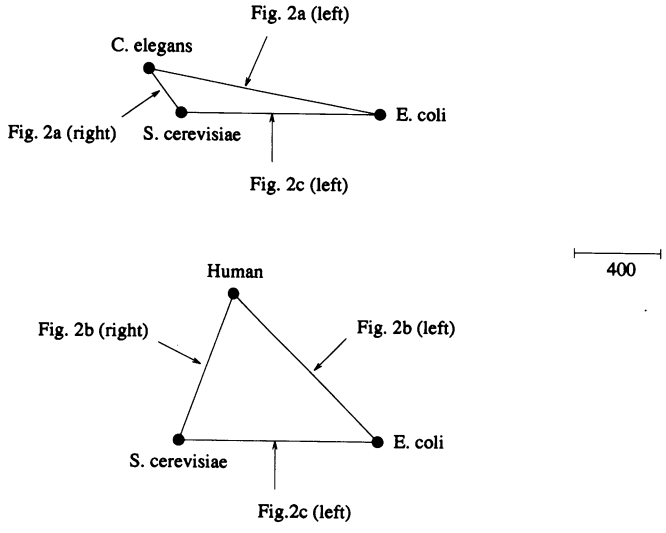

Figure 3. Pairwise distances between the control distributions shown in Fig. 2, measured at the 50% (0.50) level of the distribution function. For symmetric distributions, these distances correspond to distances between the means; this correspondence is only approximate for the asymmetric distributions in Fig. 2. The bar indicates a distance of 400 units of *Test(A,B,X)*, corresponding to the horizontal axis units in Fig. 2. Arrows show the projections corresponding to the plots shown in Fig. 2. The positions of the control sets in these plots have no significance.

E.coli and *C.elegans* control distributions, for example, is much smaller than the overlap between *S.cerevisiae* and *C.elegans* control distributions. Therefore, *E.coli* sequences, but not *S.cerevisiae* sequences, serve well as an outgroup for analyzing *C.elegans* sequences. Different sets of hexamers contribute to the overall dissimilarity between sequences from different organisms. A test sequence *X* may, therefore, be equally dissimilar to control sequences from two different outgroups *A* and *B* without implying that *A* and *B* are similar. The separations between human, yeast, and bacterial sequences shown in Fig. 2, for example, are all similar. The pairwise differences between the control sets used here are summarized in Fig. 3. It is always preferable to compare test sequences with control sets from at least two outgroups. Use of multiple outgroups also increases the likelihood of correctly identifying contaminants that are only distantly related to well-studied organisms.

Analysis of EST data sets from *Caenorhabditis* and human

We have assessed the data sets from multiple human and *C.elegans* cDNA libraries (Table 1) by comparing the sequences from each library with control sequences from the organism of origin, *E.coli* , and *S.cerevisiae*. Comparisons of 1227 ESTs (494,788 bp) from hybridization-selected (7) and 608 ESTs (189,529 bp) from unselected (8) mixed-stage *C.elegans* libraries and 714 ESTs (263,076 bp) from an embryonic *C.elegans* library (W.R.McCombie, J.C.Venter and C.Fields, in prep.) with *C.elegans*, *E.coli* , and *S.cerevisiae* control sequences are shown in Fig. 4. The ESTs from the selected library were all obtained by sequencing 5′ ends of clones, and most appear to be protein-coding sequences (7); the ESTs from the other two libraries are from both 5′ and 3′ ends. Sequences from 3′ ends contain a mixture of coding and untranslated sequences. These ESTs represent a wide range of gene families and expression classes

Table 1. A summary of abbreviation usage for the libraries in Figure 4–7

Abbreviation	library
EMB	C. elegans embryonic
Mixed	C. elegans, mixed stage
Selected	C. elegans, selected stage
HHC	Human, hippocampus
Heart	Human, heart tissue
Liver	Human, hepatoma cell line
HFB	Human, fetal brain
HIB	Human, infant brain
CCRF	Human, T-lymphoblastoid cell line

(7, 8). A distribution of *C.elegans* coding exons extracted from GenBank is also shown in Fig. 4. The cumulative distributions of the *C.elegans* EST sets all fall between the *C.elegans* genomic control and the *C.elegans* exon sequences, with the sequences from the selected library consistently the most exon-like and sequences from the embryonic library consistently the most genomic-like. The significant differences in hexamer composition between worm coding and noncoding sequences (O.White, T.Dunning and C.Fields, submitted) presumably account for at least some of the differences in the distributions between the three cDNA libraries, the exon sequences, and the genomic control. The results in Fig. 4 demonstrate the sensitivity of the hexamer composition test in comparing coding to genomic sequence data.

Laboratory *C.elegans* stocks are grown on lawns of *E.coli* ; hence, some *E.coli* contamination might be anticipated in worm cDNA libraries. The distributions of hexamer composition tests from all three libraries are well separated from the bacterial distribution; thus, these libraries show no indication of significant bacterial contamination. The worm exon and yeast control distributions overlap substantially; hence, this test could not distinguish worm sequences from yeast sequences with high confidence. Supporting the results of the hexamer composition test, no independent evidence of significant contamination with either bacterial or yeast sequences has been reported for any of these three libraries (7, 8, W.McCombie, J.C.Venter and C.Fields, in prep.).

Similar comparisons of three human brain libraries, using 1272 ESTs (394,954 bp) from a hippocampus library (1), 4749 ESTs (1,422,460 bp) from a fetal brain library (1, 2, 4), and 1862 ESTs (607,534 bp) from an infant brain library (23) with *E.coli* and *S.cerevisiae* controls are shown in Fig. 5. All three EST data sets contain 5′-untranslated, 3′-untranslated, and coding sequences; the coding sequences that have been identified by similarity searching represent a wide range of gene families, and include both highly and very weakly expressed genes (1,2,4, and 23). The HHC and HFB sets are estimated to contain roughly 30% coding sequences (4); the HIB set contains approximately 60% coding sequence (23). Less than 15% and 27% of the HHC and HFB sequences from the human libraries, respectively, have hexamer compositions that are significantly different from the control set of human sequences. As previously described (1, 2) BLAST (24) searches were conducted on sequences from the HHC and HFB sets that scored outside of the 95% upper limit

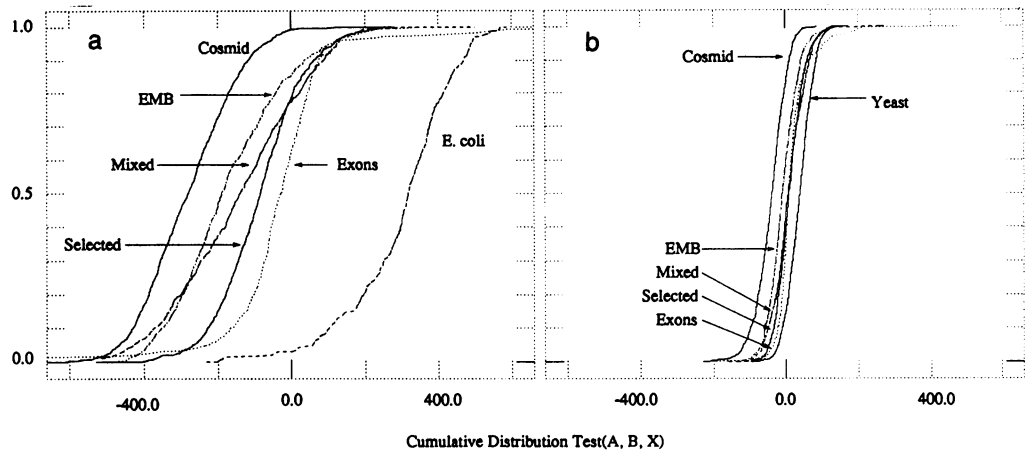

Figure 4. Comparisons of EST sequences from selected ('Selected') and unselected ('Mixed') mixed stage and embryonic ('EMB') *C.elegans* cDNA libraries with *E.coli* **(a)** and *S.cerevisiae* **(b)** sequences. *C.elegans* ('Cosmid') genomic DNA was used as the control; *C.elegans* exons extracted from GenBank were used to generate the 'Exon' distribution. The abbreviations used are summarized in Table 1. The three EST data sets consistently fall between the Exon distribution and the genomic control.

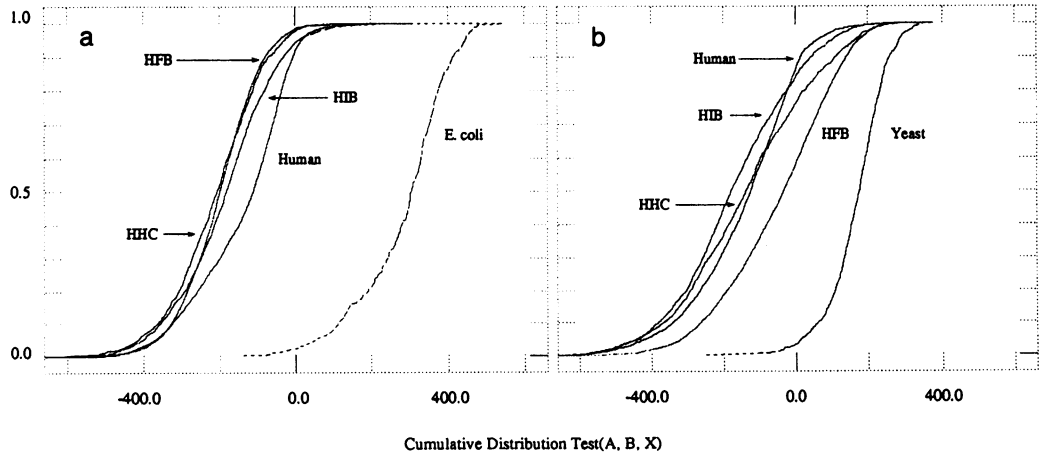

Figure 5. Comparisons of ESTs from three human brain libraries with *E.coli* **(a)** and *S.cerevisiae* **(b)** controls. HFB = fetal brain, HHC = adult hippocampus, HIB = infant brain.

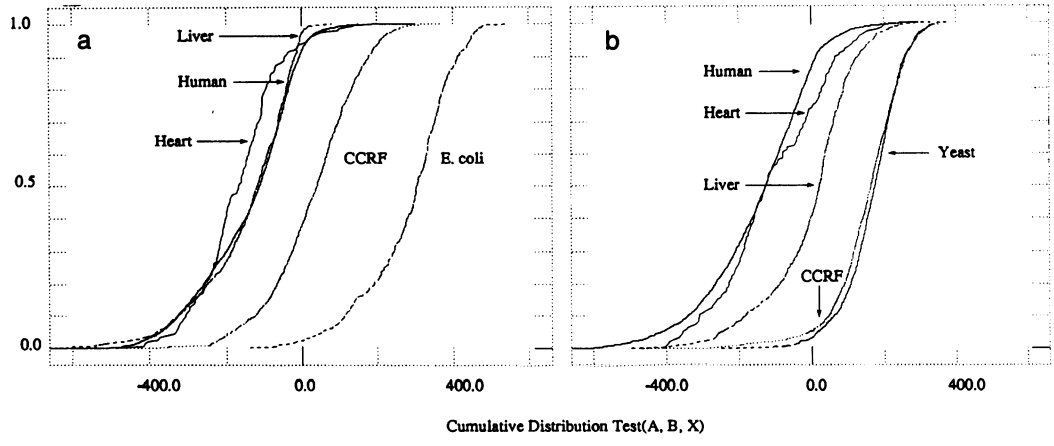

Figure 6. Comparisons of ESTs from human hepatoma('Liver'), heart, and T-lymphoblastoid libraries ('CCRF') with *E.coli* **(a)** and *S.cerevisiae* **(b)** controls. The majority of sequences from the lymphoblastoid library have a hexamer composition that is significantly different from the human control.

of variation for the human control set. Of these outlier sequences, 92% and 95% had no similarity to sequences in the data base (HHC and HFB, respectively). However, 10 HHC and 25 HFB test sequences were exact matches to human genes contained in the GenBank sequence data base. Of these exact matches to previously published sequences, 6 HHC and 13 HFB sequences contained 3 untranslated regions, one HFB sequence was from an intron, and the remaining sequences were coding regions. The exact matches to coding sequences were from a wide range of gene families. These data suggest that: a) the distributions of hexamer composition tests obtained for the three human EST sets are qualitatively similar to each other and to the human control set, indicating that the control set is reasonably representative of most sequences obtained by random selection from brain cDNA libraries; and b) up to 20% the clones from a randomly selected human cDNA library may have a hexamer composition that is not currently represented in the exon and intron sequences found in the human genes in GenBank. Those sequences that score as outliers comprise many 3′ untranslated regions and may also include a new class of coding sequences not currently represented in GenBank.

Comparisons of 631 ESTs (191,030 bp) from a human hepatoma cell line library (5), 153 ESTs (40,138 bp) from a human heart library (Strategene 936208: unpublished ESTs sequenced at the Max Planck Institute, Martinsreid obtained from the EMBL database), and 1158 ESTS (355,048 bp) from a human T-lymphoblastoid library (Clontech CCRF-CEM: unpublished ESTs sequenced at the Genethon, Paris obtained from the EMBL database) are shown in Fig. 6. The sequences from the hepatoma cell line library are principally 3′ untranslated sequences while the coding content of the heart and lymphoblastoid libraries is unknown. The distributions of the heart ESTs are similar to those of the brain ESTs shown in Fig. 5.

The distribution of the hepatoma cell line is shifted away from that of the human control sequences when compared to *S.cerevisiae* as a control. The hepatoma sequences were found to be composed of 6.3% ambiguous nucleotides (labeled 'n' in the database sequences); ESTs from heart, CCRF, and fetal brain sequences have 0.4%, 1.5% and 1.2% ambiguous nucleotides, respectively. Because the hexamer counting software does not measure hexamers that contain characters other than a, c, g, or t, the effect of ambiguous nucleotides is to reduce the total number of countable hexamers in each sequence. We have simulated the ambiguous nucleotide content of the hepatoma data by randomly adding n characters to the fetal brain data set at frequencies equal to that of the hepatoma cell line. The observed shift of the simulated data away from the human control sequences is comparable to the observed shift of the hepatoma library sequences. The presence of ambiguous nucleotides, in addition to the preponderance of 3′ untranslated sequences in this data set (5)are the most likely explanation for the shift of the hepatoma data away from the human control. There is no independent evidence for bacterial or yeast sequence matches in the hepatoma data set.

The distribution obtained from the lymphoblastoid ESTs, however, has a striking shift away from the human control sequence distribution when either *E.coli* or *S.cerevisiae* are used as controls. Consistent shifts of the entire distribution away from the human control are not observed in any of the other test sets. This suggests that the hexamer composition of the CCRF sequences are different from those in the other libraries. The shape, slope, and range of the entire lymphoblastoid hexamer

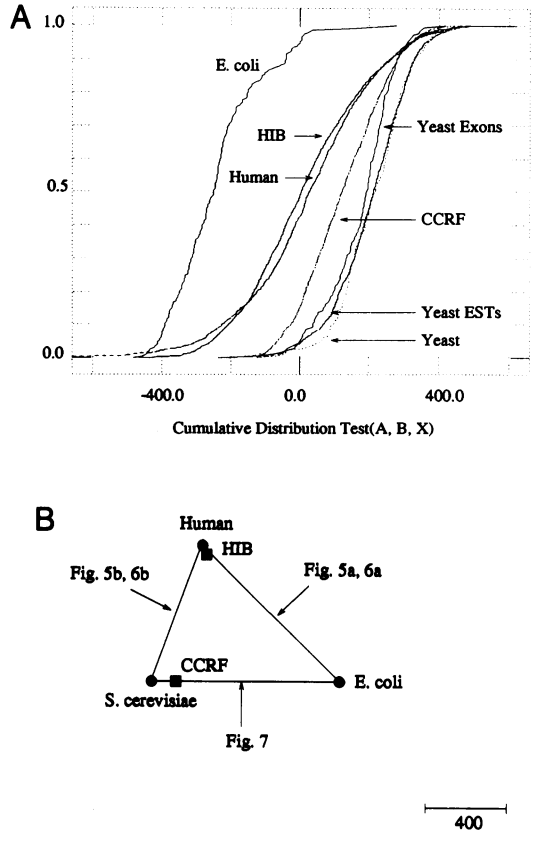

Figure 7. (a) Comparison of HIB, CCRF, and *S.cerevisiae* (K.Weinstock, M.Adams and J.C.Venter, in prep.) ESTs with *E.coli* and *S.cerevisiae* controls. Distributions for human exons and introns (combined) and yeast exons against the same controls are shown for comparison. The separation between the CCRF and *S.cerevisiae* distributions indicates that the CCRF sequences have a hexamer composition that is slightly different from the hexamer composition of *S.cerevisiae* control. (b) Pairwise distances between the data sets shown in (a), plotted as in Fig. 3. Projections corresponding to Figures 5 and 6 are indicated.

composition distribution appears similar to that of the control sequences from *S.cerevisiae* (Fig. 6b). The CCRF sequences were investigated further by calculating the cumulative distributions for these sequences compared to bacterial and yeast controls. The resulting comparison of the hexamer compositions of the CCRF library, the HIB library, a human control, a set of *S.cerevisiae* exons extracted from GenBank, and a set of 278 *S.cerevisiae* ESTs (90,199 nucleotides; K.Weinstock, M.Adams and J.C.Venter, in prep.) to yeast and bacterial controls is shown in Fig. 7. The *S.cerevisiae* exons and ESTs have distributions very similar to the yeast control. The distribution of the CCRF sequences is slightly shifted away from that of the yeast control distribution when compared to *E.coli* .

Searches using BLAST were conducted on all of the CCRF sequences. The results are summarized in Tables 2 and 3. The results of the BLAST searches support concern about the sequences in the database from this commercial library (25). Using >90% identity as the cutoff to establish the source of species for each sequence, ten CCRF clones were of human origin (0.8% from 1158 sequences); eight CCRF clones are apparently identical to nuclear human genes and two are from human mitochondria (Table 2). In the published human EST sets (Refs. 1, 2, 3 and 5) of a similar number of sequences, database searches

Table 2. BLAST best matches of >90% identity in the CCRF lymphoblastoid library

a	CCRF accession	Description	Accession of Match	%Ident
HUMAN	HUMA04C121	H. sapiens tre-2 oncogene	GBU:S112363	96.4
	HUMA88H031	H. sapiens transcription factor TCF-1 al	BB:B29911	99.4
	HUMA74A081	H. sapiens KKIALRE for ser/thre protein kinase	GB:HUMSTHPKB	98.5
	HUMA28A081	H. sapiens h-Sp1	EU:HSHSP1	97.8
	HUMA16H091	H. sapiens translationally controlled tumor protein	GB:HUMTUMP	99.2
	HUMA89C121	H. sapiens BTF3 putative transcription factor	GB:HUMBTFE	97.3
	HUMA47B051	H. sapiens Fatty-acid binding protein homologue	GB:HUMFABPHA	92.6
	HUMA23E091	H. sapiens L1 element transposable element ORFI	GB:HUMLINE1O	93.0
	HUMA26C011	Human Mitochondrial	GB:HUMMTCG	99.6
	HUMA23E041	Human Mitochondrial	GB:HUMMTCG	98.2
S. cerevisiae	HUMA05B081	S.cerevisiae Expressed Sequence Tag[b]	YAYAC71F	96.0
	HUMA70E081	S.cerevisiae Expressed Sequence Tag[b]	YAYAA65F	95.0
	HUMA04B121	S.cerevisiae retrotransposon Ty4	E:SCTY4	98.8
	HUMA08E071	S.cerevisiae centromeric region CEN11	E:SCCEN11D	98.0
	HUMA18A111	S.cerevisiae triglyceride lipase homolog	GB:S97962	97.2
	HUMA27F061	S.cerevisiae component of the pheromone signal	GBU:YSCSTE5	99.5
	HUMA30H021	S.cerevisiae CEN5-WBP1 intragenic region	GB:YSCCEN7	98.3
	HUMA39H101	S.cerevisiae spliceosome (PRP19) gene	GBU:YSCPRP19	97.1
	HUMA49B071	S.cerevisiae MPI 1 gene for mitochondrial protein	GB:YSCMPI1	100.0
	HUMA52C031	S.cerevisiae GCR2=transcriptional activator	GB:S111625 *	95.6
	HUMA61F111	S.cerevisiae Mkk2 protein kinase	GBU:YSCMKK2	96.3
	HUMA67F071	S.cerevisiae SEC6 gene	GB:YSCSEC6G	99.6
	HUMA70B031	S.cerevisiae TYE7 gene	E:SCTYE7G	95.7
L. lactis	HUMA08G021	Lactococcus lactis replication protein gene	GB:LACREPPRO	90.6
	HUMA50A081	Lactococcus lactis gene	GBU:CI5MOBPR	94.8
	HUMA58G081	Lactococcus lactis gene	GBU:CI5MOBPR	93.7
	HUMA55G061	Lactococcus lactis gene	GBU:CI5MOBPR	90.0
	HUMA62D091	Lactococcus lactis gene	GBU:CI5MOBPR	92.1
[c]	HUMA67E081	pyruvate kinase-like protein	GB:XXXPKLHO	97.8

[a]Putative origin
[b]K.Weinstock, M.Adams and J.C.Venter, in prep.
[c]Unknown bacteria, GenBank accession XXXPKLHO
Results of searching the National Center for Biotechnology Information (NCBI) nucleic acid sequence database with sequences from the CCRF-CEM lymphoblastoid library using BLAST (24). Descriptions and accession numbers of the best matches are shown for each CCRF sequence with a > 90% nucleotide similarity to a sequence in the database. BB = NCBI Backbone; GB = GenBank; GBU = GenBank Update; E = EMBL Databank; EU = EMBL Update.

would normally identify 6–11% of the sequences as exact human nuclear gene matches. The remaining matches of the CCRF sequences identified by BLAST belong to sequences from yeast, *Lactococcus lactis*, and unknown bacteria. Two CCRF clones are apparently identical to sequences from a yeast EST library (K.Weinstock, M.Adams and J.C.Venter, in prep.). The result of a peptide similarity search of 6-frame translations of the CCRF library clones with <90% identity to other published sequences is presented in Table 3. Of these, one sequence had similarity to a *Drosophila* protein and the remaining 31 matches had greatest similarity to sequences of prokaryotic origin. These data support the results for the hexamer composition test, and provide independent confirmation of the ability of the hexamer test to identify heterologous sequences from multiple, unexpected sources. They also confirm the typical result that most sequences in an EST data set are not identifiable, and hence not classifiable as to organism of origin, by BLAST.

DISCUSSION

The log-likelihood hexamer composition test described here provides a method for detecting sequences of possible heterologous origin in EST or other sequence data. This method identifies DNA sequences with hexamer compositions significantly different from that of a control sequence or set of sequences. The measure does not use a sequence alignment algorithm which requires that a sequence have a related sequence (on a nucleotide or protein level) in the public database in order to be identified. The method is independent of sequence orientation. The accuracy of the method depends on the representativeness of the control sequences that are employed and the availability of a suitable heterologous outgroup. Given a representative set of control sequences and a well-separated outgroup, the hexamer test function reliably identifies sequences that are probably not derived from the purported source organism.

Table 3. BLAST best matches of <90% identity in the CCRF lymphoblastoid library

CCRF accession	Description	Accession of Match	%Ident
HUMA37D101	Drosophila Calmodulina	GP:DROCALM4_1	48.8
HUMA80F091	Lactococcus lactis Replication protein	GP:LACREPPRO	59.7
HUMA50E061	Lactococcus lactis Transposase	GP:TRN4551AA	42.1
HUMA64H021	Lactococcus lactis Transposase	GP:LACNISC_2	39.1
HUMA01G111	Lactococcus lactis insertion sequence IS981	GP:LACNISC_2	50.0
HUMA25G031	Klebsiella terrigena Acetoin(diacetyl)reductase	GPU:KPNBUDOP	57.0
HUMA62A111	E coli SmbA	GPU:S119075_	42.6
HUMA16G041	E coli mdl gene product	GPU:ECOMDL_1	30.7
HUMA33E011	E coli mdl gene product	GPU:ECOMDL_1	42.9
HUMA09H101	E. coli DNA polymerase III catalytic subunit	GP:ECOPOLCAC	52.9
HUMA20C031	E. coli NusB - N utilization substance protein B	SP:NUSB_ECOL	70.0
HUMA43A111	Bacillus stearothermophilus Neopullulanase	PIR:A37008	28.6
HUMA09E071	Pasteurella haemolytica ORF1	GP:PASLEUTRE	53.3
HUMA10H101	Streptococcus oralis Penicillin-binding protein	GP:STRPONAA_1	58.1
HUMA42G061	Treponema pallidum Membrane protein C	PIR:A43595	47.6
HUMA46D041	Treponema pallidum Membrane protein C	PIR:A43595	48.6
HUMA23E061	Micococcus luteus Excision nuclease	SP:UVRA_MICL	41.3
HUMA30G051	N. gonorrhoeae Hypothetical protein	PIR:S19184	32.5
HUMA65D031	Listeria monocytogenes Internalin A	SP:INLA_LISM	42.5
HUMA82H111	Streptococcus matans. alpha-D galactosidase	SP:AGAL_STRM	38.4
HUMA40D041	B.fibrisolvens beta-D-xylosidase	GP:BUTXYLB_3	48.8
HUMA01C021	Streptococcus thermophilus transport system protein	SP:LACY_STRT	61.1
HUMA26B091	Pasteurella haemolytica Leukotoxin protein	SP:LKTB_PASH	47.6
HUMA01G041	Bacilus subtilis acetyl-glu-gamma-semialdehyd.	SP:ARGC_BACS	50.8
HUMA61G011	Streptococcus faecalis NADH oxidase	GPU:S114538_1	54.8
HUMA66B051	Streptococcus faecalis NADH oxidase	BB:B114539	32.4
HUMA10C041	Bacilus subtilis Phosphotransferase factor III	SP:LEVE_BACS	55.9
HUMA04A051	Plasmid pAD1 replication-associated protein	GP:AD1REPABC	43.1
HUMA31C021	Bacilus subtilis Ribsomal protein BL9	SP:RLA_BACSU	59.6
HUMA40F051	Bacilus subtilis Transcriptional activator TenA	SP:TENA_BACS	54.3
HUMA42C091	Staphylococcus aureus Transposase	SP:TNP4_STAA	39.6

Results of searching the National Center for Biotechnology Information (NCBI) peptide sequence database with 6-frame translations of sequences from the CCRF-CEM lymphoblastoid library using BLAST (24). Descriptions and accession numbers of the best matches are shown for each CCRF sequence with a < 90% nucleotide similarity to a sequence in the database for which a significant peptide database match was obtained. BB = NCBI Backbone; GP = GenPept; GPU = GenPept Update; PIR = Protein Information Resource; SP = SwissProt.

The comparisons of worm and human EST data sets shown in Figs 4–6 show that the worm and human sequences now in GenBank provide a sample that is reasonably representative, at least at the level of hexamer composition, of sequences derived from randomly selected cDNA clones. The representativeness of these control sets was confirmed by running identical tests with random selections of either 50% or 25% of the sequences in the control sets; these experiments yielded results indistinguishable from those shown in Figs. 4–6. The cumulative distributions obtained from different EST sets, such as the three *C.elegans* EST sets or the three brain EST sets analyzed here, can be significantly different; these differences appear to be due to differences in the fractions of coding sequence in the data sets. They may also reflect small systematic differences in compositions between gene families or expression classes.

The hexamer composition method described here identifies sequences that are not similar to known sequences from a putative source organism. Unlike similarity-based methods such as BLAST, this method can identify heterologous contaminants even when they are derived from previously unknown organisms. This statistical test, together with standard alignment algorithms provides a substantially more effective means of screening large sequence data sets to identify and remove heterologous sequences.

ACKNOWLEDGMENT

This work was partially supported by U.S.DOE Genome Project Grant 93ER61566 to C.Fields.

REFERENCES

1. Adams, M. D. et al. (1991) Science 252, 1651–1656.
2. Adams, M. D. et al. (1992) Nature 355, 632–634.
3. Kahn, A. S. et al. (1992) Nature Genet. 2, 180–185.

4. Adams, M. D., Kerlavage, A. R., Fields, C. and Venter, J. C. Nature Genet. 4, 256–267.
5. Okubo, K. et al. (1992) Nature Genet. 2, 173–179.
6. Hoog, C. (1991) Nucl. Acids Res. 19, 6123–6127.
7. Waterston, R. et al. (1992) Nature Genet. 1, 114–123.
8. McCombie, W. R. et al. (1992) Nature Genet. 1, 124–131.
9. Burglin, T. and Barnes T. (1992) Nature 357, 367.
10. Adams, M., Fields, C., and Venter, J. C. (1992) Nature 357, 367–368.
11. Woese, C. R. (1987) Microbiol. Rev. 51, 221–271.
12. Christen, R., Ratto, A., Baroin, A., Perasso, R., Grell, K.G., and Adoutte, A. (1991) EMBO J. 10: 499–503.
13. Volinia, S., Gambari, R., Bernardi, F., and Barrai, I. (1989) Comp. Appl. Biosci. 5:33–40.
14. Pietrokovski, S., Hirshon, J., and Trifinov, E. N. (1990) J. Biomol. Struct. Dynam. 7: 1251–1268.
15. Nussinov, R. (1991) Comp. Appl. Biosci. 7: 287–293.
16. Burge, C., Campbell, A. M., and Karlin, S. (1992) Proc. Natl. Acad. Sci. USA 89: 1358–1362.
17. Claverie, J.-M., Sauveaget I., Boubueleret I. (1990) Meth. Enzymol. 183:237–252.
18. Dunning, T. Computational Linguistics (in press).
19. Daniels, D., Plunkett, G. III, Burland, V., and Blattner, F. R. (1992) Science 257, 771–778.
20. Oliver, S. G. et al. (1992) Nature 357, 38–46.
21. Benian, G. M., Kiff, J. E., Neckelmann, N., Moerman, D. G., and Waterston, R. H. (1989) Nature 342, 45–50.
22. Sulston, J. et al. (1992) Nature 356, 37–41.
23. Adams, M., Soares, B., Kerlavage, A., Fields, C., and Venter, J.C. (1993) Nature Genet. 4 (in press).
24. Altschul, S., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) J. Mol. Biol. 215, 403–410.
25. Savakis, C. and Doelz, R. (1993) Science 259:1677.