Detection of Protein Coding Sequences Using a Mixture Model for Local Protein Amino Acid Sequence

EDWARD C. THAYER, CHRIS BYSTROFF, and DAVID BAKER

ABSTRACT

Locating protein coding regions in genomic DNA is a critical step in accessing the information generated by large scale sequencing projects. Current methods for gene detection depend on statistical measures of content differences between coding and noncoding DNA in addition to the recognition of promoters, splice sites, and other regulatory sites. Here we explore the potential value of recurrent amino acid sequence patterns 3-19 amino acids in length as a content statistic for use in gene finding approaches. A finite mixture model incorporating these patterns can partially discriminate protein sequences which have no (detectable) known homologs from randomized versions of these sequences, and from short (≤ 50 amino acids) non-coding segments extracted from the *S. cerevisiea* genome. The mixture model derived scores for a collection of human exons were not correlated with the GENSCAN scores, suggesting that the addition of our protein pattern recognition module to current gene recognition programs may improve their performance.

Key words: gene finding, mixture model, EM algorithm, sequence/structure motifs.

INTRODUCTION

DISCRIMINATION OF PROTEIN CODING SEQUENCES from noncoding sequences in genomic DNA depends on knowledge of many different sequence features which are presumably recognized by the transcription/translation machinery of the cell. As biological studies elucidate these mechanisms, computer programs incorporating the new insights will hopefully become increasingly more accurate at predicting which portions of a DNA segment encodes protein(s). Stochastic models are often used to model sequence features shown to correlate with the presence of genes. Examples of some stochastic measures, or "content statistics" as they have come to be referred, include C+G content, hexamer or dicodon usage, and DNA strand flexibility in a gene's promoter region. (See Fickett and Tung (1992) for review of different content statistics.) The presence of any one of these datum does not necessarily signify the location of a gene, but rather, the combined "additive" effect of multiple signals increases the likelihood of finding a potential gene. Guigó and Fickett (1995) demonstrated that many content statistics essentially measure little more than C+G content. Moreover, by varying the C+G content of randomly generated, nonbiological sequences, they could obtain content scores as high as any coding region they examined. One exception was the use of the Fourier transform (autocorrelation) to detect periodic usage of nucleotides over long stretches (> 100 nucleotides), and whose performance was independent of a region's C+G content.

Department of Biochemistry, University of Washington, Seattle, WA 98105.

Sequence features beyond codon and dicodon frequencies have also been used to assist gene detection. Many programs (Gish and States, 1993) utilize searches for sequence homologues to better identify and demarcate coding regions related to previously identified proteins. For example, Procrustes (Mironov *et al.*, 1998) achieved quite high accuracy in recognizing genes in human genomic sequence by searching for homologs in a database of prokaryotic proteins. While these methods make detection of homologs of previously identified genes easier, they do little to assist in the detection of genes encoding novel proteins. At the other extreme, several groups (Laub and Smith, 1998, Pachter *et al.*, 1999) have recently developed gene identification methods using nearest neighbor searches in the space of short (4 residue) peptide segments. These methods generally produce extremely high correlation coefficients, and high specificity scores and have been shown to correctly identify many exon/intron boundaries.

Starting from the assumption that recurrent local structural features in proteins should generate a finite set of recurring local patterns in protein sequences, we clustered segments of multiple sequence alignments for proteins of known structure using a measure of sequence similarity (Han and Baker, 1995, 1996). The clustering procedure yielded a large set of sequence patterns ranging from 3 to 15 residues in length, some of which correlated strongly with particular types of local structure in proteins. Subsequently, structural information was used to increase the specificity of the subset of the patterns which corresponded strongly with local structure; the resulting "I-sites library" has shown considerable promise for local protein structure prediction (Bystroff and Baker, 1997, 1998). An important feature of the sequence patterns characterized in these studies is that, unlike the patterns in the BLOCKS (Henikoff and Henikoff, 1991; Henikoff *et al.*, 1999) and PROSITE (Hofmann *et al.*, 1999; Bucher and Bairoch, 1994) databases, they are not specific for particular types of proteins; rather, they transcend protein family boundaries.

Because these sequence patterns are recurring features of protein sequences on a length scale neglected by most current methods (longer than the 4 residue peptides used by the dictionary based approaches, and shorter and more general than the matches found in the sequence homology based approaches), we reasoned that they might be useful for discriminating coding from noncoding sequences in genomic DNA. Since the database of proteins with known structure is a small subset of all known proteins, we applied the same clustering process (Han and Baker, 1995, 1996) to those Pfam (Sonnhammer and Eddy, 1997; Sonnhammer *et al.*, 1998; Bateman *et al.*, 1999) multiple sequence alignments not already accounted for by the I-sites library. In this paper we describe the development of a finite mixture model for sequence patterns which assigns a probability that a peptide segment is part of a protein sequence. We demonstrate that the model can partially distinguish coding sequence from random sequences with identical amino acid frequencies. And while the distribution of scores for noncoding sequences is clearly different from that for randomized sequences, the model performs better on coding sequences than it does on translated noncoding sequences. We compare our exon scoring method to the hexamer content statistic used by GENSCAN (Burge and Karlin, 1997) to predict exons in the Burset/Guigó gene detection test set (Burset and Guigó, 1996). In addition to showing no correlation between these two scoring methods, the comparison revealed several test set sequences which were incorrectly included in the Burset/Guigó test set due to incomplete annotation in GenBank entries, or the presence of partial genes in the original sequences.

METHODS

The likelihood ratio

$$\frac{\Pr(\text{sequence} \mid \text{coding region})}{\Pr(\text{sequence} \mid \text{non-coding region})} \tag{1}$$

is an optimal statistic for distinguishing coding from noncoding DNA. As mentioned above, most current methods replace the numerator with a statistical model typically containing a codon or dicodon measure. Since our goal was to develop a module as orthogonal as possible to current methods (to facilitate future incorporation into already existing gene finding programs), we separate the ratio into two terms

$$\frac{\Pr(\text{sequence} \mid \text{coding region})}{\Pr(\text{sequence} \mid \text{non-coding region})} = \frac{\Pr(\text{sequence} \mid \text{amino acid frequencies in proteins})}{\Pr(\text{sequence} \mid \text{non-coding region})}$$
$$\times \frac{\Pr(\text{sequence} \mid \text{coding region})}{\Pr(\text{sequence} \mid \text{amino acid frequencies in proteins})}. \tag{2}$$

The first term is sensitive to amino acid (codon) composition, and the second, to the ordering of the amino acids. Here we focus entirely on the second term.

The model

Imagining the space of all biological sequences of length L to consist of M different subpopulations, it seems reasonable to decompose Pr(sequence | coding region) into the separate probabilities of observing the sequence amongst any one of the subpopulations, properly accounting for the size of the subpopulations. Such a process is described by a mixture model containing M components $\{C_1, C_2, \ldots, C_M\}$, which has the form

$$\Pr(x \mid \text{coding region}) = \sum_{i=1}^{M} \Pr(C_i) \Pr(x|C_i) \tag{3}$$

where x is a peptide of length L. We make the simplifying assumption of positional independence amongst the amino acids along x, and assume the probability of an amino acid at any given position in a component C_i is given by a multinomial distribution. These assumptions are commonly made when one is modelling protein families based on multiple sequence alignments of the family members (Hertz *et al.*, 1990; Bailey, 1993; Sjolander, 1997; to mention only a few). Hence for a single sequence x of length L one has,

$$\Pr(x|C_i) = \prod_{l=1}^{L} \Pr(x[l]|C_i, \ \text{position } l), \tag{4}$$

where $x[l]$ is the lth amino acid along x.

The parameters describing a component C_i are its frequency table (sometimes called a *profile* or *weight matrix*),

$$\{\Pr(aa|C_i, \ \text{position } l)\}, \tag{5}$$

and an a priori probability or "mixing coefficient" $\Pr(C_i)$, which estimates the fraction of the space of all biological sequences that are characterized by C_i. The a priori probabilities are meant to appropriately weight the contribution of each cluster's probability score to the combined model score, and represent an added statistic beyond the frequency tables for individual clusters. Since we will only be interested in the log likelihood ratio

$$\log\left(\frac{\Pr(x|C_i)}{\Pr(x|B)}\right), \tag{6}$$

where B is the background model—also assumed to be a multinomial distribution with amino acid emmision probabilities equal to the genome wide amino acid frequencies—it is sufficient to consider

$$\log\left(\frac{\Pr(x|C_i)}{\Pr(x|B)}\right) = \sum_{l=1}^{L} \log\left(\frac{\Pr(x[l]|C_i, \ \text{position } l)}{\Pr(x[l]|B, \ \text{position } l)}\right). \tag{7}$$

Determining the appropriate number of model components M was done empirically, and three models are discussed in which $M = 76$, $M = 676$, and $M = 2076$. We chose $L = 19$ for this study based on previous studies of the correlation between local sequence and local structure (Bystroff and Baker, 1998; Han and Baker, 1996, 1995; Salamov and Solovyev, 1995; Yi and Lander, 1993).

Model component profiles

I-sites derived components. We constructed our initial model, referred to as ISL, using the I-sites (Bystroff and Baker, 1998) cluster profiles. These cluster profiles were previously shown to correlate with local, three dimensional structural patterns. Detailed descriptions of each I-sites cluster are given by Bystroff and Baker (1997, 1998). Since the I-sites profiles varied in length from 3-15 residues, they were extended to 19 residues to simplify comparison of $\Pr(x|C_i)$ for different clusters. Placing the initial profile centrally inside the 19 residue profile, we used the I-sites training set and cluster membership lists to measure the observed amino acid frequencies for the new flanking positions of each profile.

Pfam derived components. There are additional sequence patterns which are detectable but which do not correlate strongly with local structure (Han and Baker, 1996) or for which there is no known three dimensional structural data. To investigate the potential of such patterns to detect genes we extended the ISL model to include components derived from cluster analysis (KMeans (Duda and Hart, 1970)) of that portion of the training set which the ISL model does not attribute a high probability. Several metrics for clustering were tested, and symmetrized relative entropy (Sjolander, 1997)

$$D(x_j[l], C_i[l]) = \sum_{aa=1}^{20} \left[\Pr(aa|x_j[l]) - \Pr(aa|C_i[l]) \right] \log \left(\frac{\Pr(aa|x_j[l])}{\Pr(aa|C_i[l])} \right) \tag{8}$$

gave consistently better clusters as measured with the independent test sets described below. Specifically, we scored the training set with the ISL model and removed any segments with a log likelihood ratio greater than a preselected value. Segments scoring higher than the cutoff were considered to be accounted for by the ISL derived components of the model. This resulted in approximately 120,000 segments which were clustered into either $K = 600$ or $K = 2000$ clusters. All the cluster centers were taken as initial estimates for class profiles and were added to the ISL model to create the ISL+600 and ISL+2000 models. Several clusters from each model have been found to correlate strongly with new local structures which were poorly represented in the initial I-sites study (Bystroff and Baker, 1997, 1998) (data not shown).

Background mixture component. Since we expect the model components to be exhaustive, i.e., $\sum_{i=1}^{M} \Pr(C_i) = 1$, *all* models include a "background" component whose profile is identical at all residue positions and whose expected amino acid frequencies are equal to the overall amino acid frequencies of the sequence dataset being evaluated. This component's prior probability indicates what portion of the training set is still unaccounted for by the other model components.

Model component priors

Unlike the class profiles, the class priors $\Pr(C_m)$, which depend intimately on the training set used, were completely unknown and were optimized using the EM algorithm. So as to access potential convergence to local critical points in the maximization process, several different initial estimates were tried for these class priors.

Training the model

To make use of the large number of distinct sequences available for some protein families without allowing the model to be dominated by large families, we used multiple sequence alignments rather than single sequences as the training set. The training set consisted of multiple sequence alignment segments taken from the full family alignments from the Pfam 2.1 database (Sonnhammer and Eddy, 1997; Sonnhammer *et al.*, 1998; Bateman *et al.*, 1999). Only those positions which contained more than 50% nongap characters from families containing at least 15 aligned sequences were initially considered. This set of segments occasionally contained sequences which were considered to be low complexity by the filter SEG (Wooton and Federhen, 1993). Low complexity sequences clearly invalidate many of the statistical assumptions commonly made about sequences, in particular, positional independence is so badly violated that the prior probability parameter $P(C_i)$ for a component attempting to model a collection of low complexity sequences becomes exaggeratedly large due to the fact that each sequence is counted approximately L times rather than once. To alleviate the influence of these sequences, the consensus sequence for each multiple sequence alignment meeting the above criteria was evaluated with SEG and those positions which were not marked as low complexity were used. This resulted in a training set containing $\approx 250,000$ residues with $\approx 125,000$ trainable 19 residue segments. For each of these positions the observed amino acids were converted to amino acid frequencies using Sjolander's nine-component Dirichlet prior model (Sjolander *et al.*, 1996; Sjolander, 1997).

Given a model component C_i, and a collection $\{m_t\}$ of multiple sequence alignment segments thought to belong to C_i, the maximum likelihood estimator for the parameter $\Pr(aa|C_i$ position $l)$ is

$$\frac{\sum_t n_t(aa)}{\sum_{aa} \sum_t n_t(aa)} \tag{9}$$

where $n_t(aa)$ is the number of times amino acid aa appears in the multiple sequence alignment m_t (properly weighted to account for biased sampling in the alignment) at position l. If one assumes that each multiple sequence alignment contributes an equal number of "counts" to the component C_i, i.e., $N_t = \sum_{aa} n_t(aa)$ is constant for all t, this expression simplifies to the average frequency of amino acid aa at position l of the collection of multiple sequence alignments, i.e.,

$$\frac{1}{T} \sum_{t=1}^{T} \frac{n_t(aa)}{N_t}. \tag{10}$$

Estimation of the model's mixing coefficients $\{\Pr(C_i)|i = 1, 2, 3, \ldots, M\}$ was done using the Expectation-Maximization (EM) algorithm to maximize the log likelihood of observing the training set given model (3) with respect to the parameters $\Pr(C_m)$. Specifically, we maximized

$$\sum_{j=1}^{J} \log \left(\sum_{m=1}^{M} \Pr(C_m) \prod_{l=1}^{L} \prod_{aa=1}^{20} \left[\frac{\Pr(aa|C_m, \text{ position } l)}{\Pr(aa|B)} \right]^{n(aa|x_j[l])} \right), \tag{11}$$

where J is the number of 19 residue multiple sequence alignment segments x_j in the training set, and $n(aa|x_j[l])$ is a normalized count of the number of amino acids aa found at position l of x_j.

Test sets

All test sets contain amino acid sequences and every possible 19 residue segment is considered, i.e., a 19 residue long window is shifted by one along the length of the translated peptide segments described below.

Our first test set consisted of those Pfam families which contained only two sequences at least one of which was human. We chose the longer of the two sequences. This test set is referred to as the **HPfamB** test set and contains 925 proteins. The test set sequences are sufficiently different from all other protein sequences that they were not included in the other families of Pfam (Sonnhammer and Eddy, 1997). By construction these sequences are not present in the training set.

The availability of the genomic sequence of *S. cerevisiae* and various results on the classification of the open reading frames from the genome has provided an excellent source for test sequences. We have extracted three different test sets from these resources.

1) **YCIV** was constructed at the Munich Information Centre for Protein Sequences (MIPS) (MIPS, 1997) and contains 786 ORFs similar to a previously identified protein of unknown function (obtained file 7/97). There are 341,377 scorable 19 residue segments in this set.

2) **YCV** was also constructed at MIPS (MIPS, 1997) and contains 1063 ORFs which were found to have no similarities to any previously identified proteins (obtained file 7/97). There are a total of 363,684 scorable 19 residue segments in this set.

3) **YCNC** contains a collection of 19 to 40 amino acid long segments translated from regions of the yeast genome lying outside the set of all open reading frames with length greater than 50 amino acids. Specifically, each chromosome's sequence was masked with RepeatMasker (Smit and Green, 1997) for centromere sequences, and transposable elements, then all ORFs with length greater than 50 amino acids were removed. From the remaining DNA, we extracted any segment between 19 and 40 amino acids in length that was flanked by stop codons and contained no start codon. All segments were mutually exclusive. YCNC contains 49,621 testable 19 residue segments.

Nothing in the process which generates YCNC identifies pseudogenes and removes them, so a segment meeting our selection criteria may have been taken from inside a pseudogene (two YCNC segments were found opposite the repeative sequences at the 3' end of the flocculation pseudogene on chromosome I).

YCNC serves as a negative control in our testing proceedure. Recall that our models were not designed to optimally distinguish between segments from proteins and segments from noncoding DNA, so that any discrimination achieved results simply from the lack of protein like sequences in these noncoding segments.

In addition to the test sets described above, we also considered three collections of exons from the Burset/Guigó gene finding test set (Burset and Guigó, 1996). Burge provided the correctly predicted, incorrectly predicted, and missed exons generated from GENSCAN (Burge and Karlin, 1997) for this commonly used test set. Each exon was accompanied by its GENSCAN content measure. In order to compare these full length exon scores to scores generated from our best model, we needed to create a full length exon score. This was accomplished by summing all 19 residue model scores across the length of the exon, and then converting this sum to a z-score based on the summed scores of 100 randomized versions of the same exon.

RESULTS AND DISCUSSION

Each mixture model was tested to determine how well it distinguishes protein segments from randomly shuffled amino acid sequences. Each of the four test sets (YCIV, YCV, YCNC, HPfamB) were scored independently, and the frequency of each score value is shown in Figure 1.

It is important to remember that even though the majority of our test sets contain yeast sequences, the models are not trained solely on yeast proteins. The training set contains protein sequences from a wide variety of organisms both prokaryotic and eukaryotic.

For any 19 residue segment x from a test sequence, we compute the log likelihood ratio

$$\log \left(\frac{\Pr(x\,|\text{model})}{\Pr(x\,|B)} \right) \tag{12}$$

and bin these values to give the score distributions shown in Figure 1. To determine a background score distribution for each test set which reflects its amino acid composition, we also shuffle each test set sequence and subject these "randomized" sequences to the same scoring process. Each test set's background score distribution is shown by a solid line in Figure 1. The model performed similarly on each randomized set, suggesting it is more sensitive to the order of the amino acids than the amino acid composition of the sets.

The larger of the three models, ISL+2000, out performs the other two models. On average, the clusters from the ISL+2000 model contain 60 multiple sequence alignment segments resulting in a higher specificity than either of the other two models. Increasing the specificity of the individual model components does not in general increase the model's overall performance as was discovered when we tested a model containing 5000 components constructed in an identical fashion to the models presented. The 5000 component model performed similarly to the original I-sites model, probably demonstrating an over training phenomenon. As more proteins are added to the training database the optimal number of model components may change depending on whether these additions contribute detectable new sequence patterns. It often takes several examples of a given pattern before it will be found with the techniques employed.

Clearly the yeast noncoding sequences do not behave like their randomized counterparts despite having the same amino acid composition. This may reflect the presence of pseudo genes, unmasked repeat elements not yet represented in the repeat element database, and low complexity nucleotide sequences. Three noncoding segments were found to be taken from the flocculation pseudo gene in chromosome I, suggesting the possibility of other pseudo gene derived sequences being present.

Table 1 contains the fraction of all 19 residue segments from each test set which scored above a log likelihood ratio (LLR) of 1 when one discounts the observed number of segments by the expected number as measured by the score distribution for the randomized sequences. This corresponds to the area between the randomized and nonrandomized score curves of Figure 1 above the score of 1 LLR normalized by the number of segments scored in each test set. Pfam is continually being updated and with each new release we have tried to rerun the KMeans clustering and model training with the updated training set. Pfam release 3.3 doubled the size of the training set, but a 2000-component model trained on this larger dataset performed insignificantly better on all test sets.

Figure 2 depicts the locations of all 19 residue long segments scoring better than 3 LLR with the ISL+2000 model on the first 500 proteins from the HPfamB test set, demonstrating that the difference between randomized peptides and nonrandomized proteins is not due solely to a select set of regions on a few proteins. It is clear from these images that in many cases the differences between randomized and nonrandomized proteins can be seen at the individual protein level. Moreover, despite occasional high
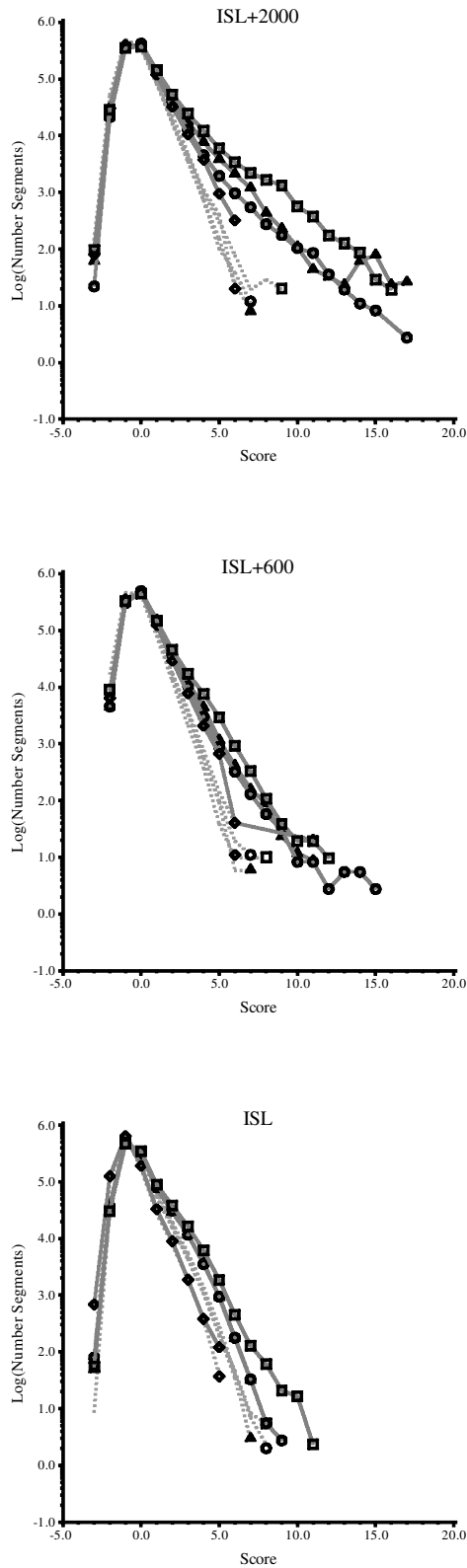
FIG. 1. Mixture model derived log likelihood ratio score distributions. Each 19 residue long segment from each test set (YCIV – triangles, YCV – circles, YCNC – diamonds, HPfamB – squares) is scored with each mixture model (ISL – bottom, ISL+600 – middle, ISL+2000 – top) and the frequency of each binned score is computed. Additionally, a score distribution for each test set's randomization is computed (shown with dotted lines). Each histogram is renormalized to contain 100,000 scored segments for comparison purposes.

TABLE 1. PERCENTAGE DIFFERENCES BETWEEN SCORE
DISTRIBUTION AND RANDOMIZED TEST SET'S SCORE
DISTRIBUTION FOR ALL SEGMENTS SCORING
BETTER THAN 1 LLR

	ISL	ISL+600	ISL+2000
YCIV	4.8%	10.7%	11.1%
YCV	5.2%	9.3%	9.0%
YCNC	0.6%	4.0%	4.0%
HPfamB	6.0%	12.6%	14.3%

scoring segments in the randomized proteins, these segments are seldom surrounded by other high scoring segments, as they are in the nonrandomized proteins.

Each correctly predicted exon from the GENSCAN exons was scored with the ISL+2000 model using the full length scoring method described above. These ISL+2000 z-scores were compared to the GENSCAN content scores to determine if these two potentially independent scoring functions were strongly correlated. A weak correlation might suggest that the addition of our model scores to GENSCAN (and potentially other gene finding programs) would improve the gene detection performance. The regression coefficient between these two scores was $R = 0.28$, supporting the proposal that our models are measuring features different from the hexamer or dicodon statistic.

Any GENSCAN missed or incorrectly predicted Burset/Guigó exon which scored greater than 2.0 ISL+2000 z-scores was examined. Several of the incorrectly predicted exons belong to a duplicated γ-globin-1 gene which fails to be expressed (Fitch *et al.*, 1991). The annotation (GenBank Accession X53419) for this gene contains a **TATA_signal**, three exons, and a **PolyA_signal**, but lacks a **CDS** or **mRNA** entry which is a required keyword in order for the Burset/Guigó test set to recognize a gene in the sequence. Because of the missing CDS field, these exons are defined as false exons for the test.

Several groups have recently developed gene identification methods using nearest neighbor searches in the space of short peptide segments. These methods generally produce extremely high correlation coefficients, and high specificity scores and have been shown to correctly identify many exon/intron boundaries. These approaches are very similar to our mixture model in that both should allow for the detection of novel genes comprised of small re-usable protein subunits. It is instructive to compare our approach to the dictionary based methods (Laub and Smith, 1998; Pachter *et al.*, 1999) mentioned in the introduction. An obvious difference is the length scale of the sequence patterns: 19 residues in our model, and 4 residues in the dictionary based approaches. More generally, the approaches differ from each other in a way that has classically divided researchers in the field of classification theory. The dictionary, or "word look-up," methods search the space of known protein segments for the nearest segment to the candidate protein segment. The closer the nearest neighbor is to the candidate segment, the higher the probability that the candidate is coding for a protein. In contrast, the mixture model approach starts by first finding highly populated regions of the space of all protein segments, and then developing statistical models for these neighborhoods. In general, use of a statistical model rather than a "look up" approach loses some specificity, but gains generalizability, since a never before observed protein segment may be well described by one of the component distributions. It is worth noting that the mixture model approach ultimately converges to a dictionary style method as the number of components increases to the number of protein family sub-segments present in the training set (ignoring the differences between using multiple sequence alignments of families instead of individual family members). Both methods seem worthy of futher development and evaluation, and in a time when so much raw genomic sequence is being generated, there is a clear need for multiple approaches to solving the gene detection problem.

CONCLUSION

The identification of potential genes in genomic sequence has become an integral part of the analysis of genomic sequence, and while not being the conclusive proof of the existence of a gene, such predictions often lead researchers to interesting regions for further analysis.

HPfamB Sequences   Randomized HPfamB Sequences
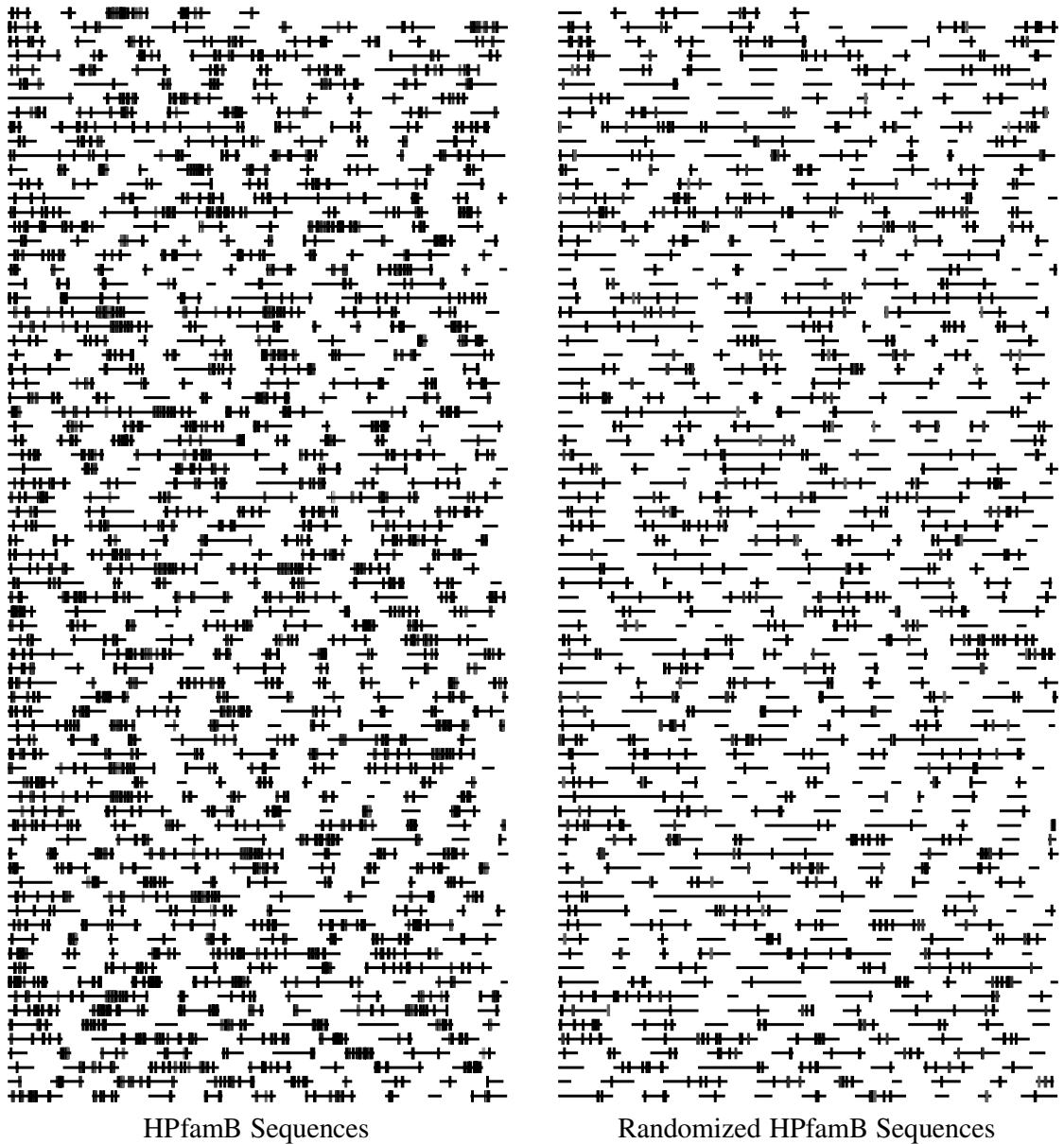
FIG. 2. ISL+2000 3 LLRs or better hits on the first 600 proteins from the human PfamB test set. Each protein is represented by a horizontal line with length scale adjusted to accommodate the longest protein in the specified display space. Each vertical bar or tick mark represents the 19 residue long segment scoring above 3 LLRs, grey scale coded with lighter shades corresponding to higher LLR scores. The image is divided into two panels. The left panel depicts the ISL+2000 model hits, and the right panel depicts the ISL+2000 model hits in one randomization of the same proteins. For comparison purposes, each protein is in the same location in the two panels.

We have described a mixture model based on sequence patterns 19 residues in length which complements and extends the content statistics used in current gene finding programs. The three protein sequence test sets (YCIV, YCV, and HPfamB) demonstrate that our models partially discriminate between some protein segments and random peptides with the same global amino acid frequencies, and, though to a lesser degree, between protein sequences and translated noncoding yeast sequences.

On what length scale do protein sequences differ from random sequences with the same amino acid composition? The differences in the score distributions of our models on the coding and scrambled coding sequence sets (Figure 1) demonstrates that there are clear nonrandom features in protein sequences on a nineteen residue length scale. Models based on ten residue segments (data not shown) also partially discriminate coding sequence from scrambled coding sequence, suggesting there are nonrandom features

already at the ten residue length scale. The origin of these nonrandom features is likely to be due, at least in part, to the distinct sequence preferences of common local structural motifs from which protein tertiary structures are built (Bystroff and Baker, 1998).

The models described here appear to capture features of protein sequences not incorporated into current gene finding methods. The correlation coefficient between our full length exon scores and the hexamer like statistic computed in GENSCAN for the Burset/Guigó test set was quite low (0.28). We are eager to assist in the incorporation of our models into existing gene detection programs.

ACKNOWLEDGMENTS

REFERENCES

Bailey, T.L. 1993. Likelihood vs. information in aligning biopolymer sequences. Technical Report CS93-318, Univ. California, San Diego.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.*, 27(1):260–262.

Bucher, P., and Bairoch, A. 1994. A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation. *Proceedings 2nd International Conference on Intelligent Systems for Molecular Biology*, 53–61.

Burge, C., and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268, 78–94.

Burset, M., and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics*, 34, 353–367.

Bystroff, C., and Baker, D. 1997. Blind predictions of local protein structure in CASP2 targets using the I-sites library. *Proteins*, Suppl 1, 167–171.

Bystroff, C., and Baker, D. 1998. Prediction of local protein structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, 281(3), 565–577.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Duda, R.O., and Hart, P.E. 1970. *Pattern Classification and Scene Analysis* California Artificial Intelligence Group, Stanford Research Institute, Menlo Park, CA.

Fickett, J.W., and Tung, C.-S. 1992. Assessment of protein coding measures. *Nucl. Acids Res.*, 20(24), 6441–6450.

Fitch, D.H., Bailey, W.J., Tagle, D.A., Goodman, M., Sieu, L., and Slightom, J.L. 1991. Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proc. Natl. Acad. Sci. U.S.A.*, 88, 7396–7400.

Gish, W., and States, D.J. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* 3, 266–272.

Guigó, R., and Fickett, J.W. 1995. Distinctive sequence features in protein coding genic non-coding, and intergenic human DNA. *J. Mol. Biol.*, 253, 51–60.

Han, K.F., and Baker, D. 1995. Recurring local sequence motifs in proteins. *J. Mol. Biol.*, 251, 176–187.

Han, K.F., and Baker, D. 1996. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci. USA, Biophysics* 93, 5814–5818.

Han, K.F., Bystroff, C., and Baker, D. 1997. Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Science*, 6(7), 1587–1590.

Henikoff, S., and Henikoff, J.G. 1991. Automated assembly of protein blocks for database searching. *Nucl. Acids Res.*, 19(23), 6565–6572.

Henikoff, S., Henikoff, J.G., and Pietrokovski, S. 1999. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, 15(6), 471–479.

Hertz III, G.Z., Hartzwell, G.W., and Stormo, G.D. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comp. Appl. in Bios.*, 6(2).

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. 1999. The PROSITE database, its status in 1999. *Nucl. Acids Res.*, 27(1), 215–219.

Laub, M.T., and Smith, D.W. 1998. Finding Intron/Exon Splice Junctions Using INFO, INterruption Finder and Organizer *Journal of Comp. Bio.*, 5(2), 307–321.

MIPS. 1997. URL `http://muntjac.mips.biochem.mpg.de/ycd/classes/classification.html`

Mironov, A.A, Roytberg, M.A., Pevzner, P.A., and Gelfand, M.S. 1998. Performance-Guarantee Gene Predictions via Spliced Alignment. *Genomics*, 51, 332–339.

Pachter, L., Batzoglou, S., Spitovsky, V.I., Beebee Jr., W.S., Lander, E.S., Berger, B., and Kleitman, D.J. 1999. A Dictionary Based Approach for Gene Annotation. *J. Comp. Bio.*, 6(4), 491–430.

Salamov, A.A., and Solovyev, V.V. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.*, 247, 11–15.

Sjölander, K. 1997. *A Bayesian-information theoretic method for evolutionary inference in proteins.* PhD thesis, Univ. California, Santa Cruz.

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, 12(4), 327–345.

Smit, A.F.A., and Green, P. Repeatmasker. Used a local copy of Repeatmasker with a yeast specific database of repeats and elements constructed by A. Smit. For vertebrate sequence, Repeatmasker is available on the WWW at URL `http://ftp.genome.washington.edu/RM/RepeatMasker.html`

Sonnhammer, E.L.L., and Eddy, S.R. 1997. Pfam: A comprehensive database of protein families based on seed alignments. *Proteins*, 28(3), 405–420.

Sonnhammer, E.L.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl. Acids Res.*, 26(1), 320–322.

Wooton, J.C., and Federhen, S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chemistry*, 17, 149–163.

Yi T.-M., and Lander, E.S. 1993. Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, 232, 1117–1129.

Address correspondence to:
Edward C. Thayer
1201 Eastlake Ave. E.
Seattle, WA 98102