Locating protein-coding regions in human DNA sequences by a multiple sensor–neural network approach

(coding exon localization/gene structure/pattern recognition/DNA sequence analysis)

EDWARD C. UBERBACHER*† and RICHARD J. MURAL*

*Biology and †Engineering Physics and Mathematics Divisions, Oak Ridge National Laboratory, and the University of Tennessee–Oak Ridge Graduate School of Biomedical Sciences, Oak Ridge, TN 37831

Communicated by William L. Russell, August 21, 1991 (received for review May 24, 1991)

ABSTRACT Genes in higher eukaryotes may span tens or hundreds of kilobases with the protein-coding regions accounting for only a few percent of the total sequence. Identifying genes within large regions of uncharacterized DNA is a difficult undertaking and is currently the focus of many research efforts. We describe a reliable computational approach for locating protein-coding portions of genes in anonymous DNA sequence. Using a concept suggested by robotic environmental sensing, our method combines a set of sensor algorithms and a neural network to localize the coding regions. Several algorithms that report local characteristics of the DNA sequence, and therefore act as sensors, are also described. In its current configuration the "coding recognition module" identifies 90% of coding exons of length 100 bases or greater with less than one false positive coding exon indicated per five coding exons indicated. This is significantly lower false positive rate than any method of which we are aware. This module demonstrates a method with general applicability to sequence-pattern recognition problems and is available for current research efforts.

Major sequencing efforts, a number of which are directed toward identifying and localizing human disease genes (1–3), are producing large amounts of DNA sequence data. DNA sequencing is no longer restricted to regions of known function, and large amounts of anonymous DNA sequence are being generated. Indeed the use of sequence-tagged sites (4) to mark positions on a high-resolution map of the human genome will result in the generation of many anonymous stretches of DNA sequence.

Perhaps the most fundamental question that can be asked about a DNA sequence is whether or not it encodes protein. Recognition of coding sequence has generally been approached by examining the positional and compositional biases imposed on the DNA sequence in protein-coding regions by the genetic code and by the distribution of amino acids in proteins (5–7). While it is fairly easy to devise algorithms with some value as predictors of coding regions, such methods are generally not sufficiently robust to be useful for finding genes in anonymous DNA sequence data from higher eukaryotes. Rather than attempting to develop a single "perfect" indicator of protein-coding character, given the incomplete state of our understanding of the underlying biological processes, a more useful approach may be to consider how information from several algorithms, each designed to recognize particular sequence properties, can be combined to provide more powerful pattern recognition capabilities (8, 9).

We regard the recognition of patterns in DNA sequence data as a problem analogous to environmental sensing in sensor-based robotic systems. In these systems, perception of the robot’s surroundings occurs through the integration of information from multiple sensors—e.g., charge-coupled device (CCD) cameras, sonar transducers, laser range finders, tactile sensors (9). Optimal integration of the outputs from these sensors, accomplished through machine learning, allows one to form a combined best estimate of the environment that is better than any estimate based on individual sensors.

We have applied this approach to DNA sequence analysis and particularly to the problem of recognizing coding sequences. This "coding recognition module" (CRM) (Fig. 1) incorporates a group of seven sensor algorithms (described below) each designed to provide an indication of the coding potential of a region of sequence. A neural network is used to integrate the sensor outputs and to predict the location of coding regions. After a suitable training procedure, the neural network learns how to interpret the sensor outputs, and, when provided with sensor data from a test sequence, can make accurate decisions about the location of coding DNA.

There are a number of advantages to the DNA pattern recognition method implemented in the CRM. (i) It represents a convenient framework for combining different types of input information (for example, statistical information and rules). (ii) The various types of input information are integrated, in a nonbiased manner, by machine learning algorithms (i.e., neural-network training). (iii) The system is robust to input (sequencing) errors because of the redundant and partially independent nature of the input sensors. The modular nature of this system also allows it to evolve by incorporating new sensor algorithms as they become available.

Several of the sensors measure characteristics of DNA sequences that are organism-specific (10). Although preliminary results suggest that the CRM based on statistics compiled from human DNA sequences performs well on sequences from other mammals, it is relatively simple to create similar modules for other species.

Sensors

To determine the likelihood that a given sequence position is within a coding segment, a program written in C language calculates the values of the seven sensors for a 99-base sequence window centered at each test position in the sequence, and the sensor signals, scaled between 0.0 and 1.0, are then evaluated by the neural net. After neural-network training is complete, the weights of the net are extracted and incorporated into a C routine that simulates the action of the net. A brief overview of the sensor algorithms follows.

(i) Frame bias matrix. The basis of this method is the nonrandom frequency with which each of the four bases occupies each of the three positions within codons. This distribution is due to unequal usage of amino acids and to preferred use of codons for particular amino acids (10). This

Abbreviation: CRM, coding recognition module.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.
bias, expressed as a matrix, is used as a probe to identify potential coding regions and the preferred reading frame. If a region codes for protein, then one frame should have a significantly better correlation to the bias matrix than the other two possible reading frames. The correlation coefficient between the matrix and each reading frame is calculated and the difference between the best and worst coefficient is used as an indicator of coding potential.

(ii) Fickett. This is an implementation of an algorithm developed by Fickett (5) that considers several properties of coding sequences. In a given window, it independently examines the 3-periodicity of each of the four bases and compares them to the periodic properties of coding DNA. It also compares the overall base composition of the test DNA with the known composition for coding and noncoding DNA.

(iii) Dinucleotide fractal dimension. Dinucleotide occurrence is known to be far from random, with dinucleotides such as AA and TC being common and CG being rare. By examining the transitions of sequential dinucleotides (asking whether the next dinucleotide is of similar or different commonality), it is possible to view a DNA sequence as a dynamic function. These transitions can be considered as changes in energy, in the Boltzmann sense, by using the energy scale $E = -\ln(p)$, where $p$ is each dinucleotide's probability. These fluctuations can be characterized by a fractal dimension (11). We have found that coding DNA usually has lower dimension than noncoding DNA (9). The sensor value represents the difference in fractal dimension between a reference value derived from intron DNA and the dimension of the window being examined.

(iv) Coding 6-tuple word preferences. One way of characterizing sequences is by examining the frequency of occurrence of nucleotide "words" of a given length in the sequence. Different types of DNA sequence (introns, coding regions, etc.) have different distributions of word occurrence (19). Each word's preference value is calculated as the logarithmic ratio of its normalized frequency of occurrence in coding vs. noncoding human DNA, and the sum of preference values in the window provides the coding indicator. The 6-tuple frequencies for protein-coding DNA were compiled from the protein-coding portions of 122 human cDNA sequences (about 210,000 nucleotides), and the statistics for noncoding DNA were compiled from a data set of about 175,000 bases of sequence from human introns. These sequences were extracted from GenBank release 60.0 (12).

(v) Coding 6-tuple in-frame preferences. This is similar to the previous sensor except that the observed 6-tuples in the test DNA are compared with the preference values of in-frame 6-tuples compiled from coding DNA. The total preference is computed three times for the analysis window, once for each reading frame. The predicted reading frame is taken to be the one that provides the best 6-tuple in-frame coding vs. noncoding preference, and the sensor value corresponds to the total preference for this frame.

(vi) Word commonality. The overall frequency of occurrence or commonality of a given 6-tuple in bulk DNA is related to its context. For example, introns use extremely common words and exons relatively rare words. In this sensor, 6-tuple word commonality is defined as the logarithmic ratio of its normalized frequency of occurrence (in a subset of human genomic DNA of about 900,000 bases) divided by its expected random frequency. The score at each test position for 6-tuple commonality is calculated by summing all (overlapping) 6-tuple commonalities contained completely in the analysis window.

(vii) Repetitive 6-tuple word preferences. The test sequence is compared with 6-tuple statistics for several classes of repetitive DNAs in a manner similar to that used in the previous three algorithms, and the largest total preference in the window (best similarity to a repetitive type) is used as the sensor. This is a negative coding indicator since it reflects the fact that highly repetitive DNA rarely encodes protein.

As an example of the application of these seven algorithms, Fig. 2 A-G shows the sensor outputs for the 6500-base-pair human ras protooncogene region.

Neural Network

A back-propagation neural network (13, 14) was constructed to integrate the output from the seven sensor algorithms. This network consists of 7 input nodes, two hidden layers of 14 and 5 nodes, and an output node. Input to the neural network for both training and test sets consists of a vector containing the values of the seven sensor algorithms (Fig. 2 A-G) calculated for positions at intervals of 10 bases along the sequences of...
interest.† For training, the net is also provided with the proper output value (1 for coding, 0 for noncoding) for each position. During learning the network compares its estimate of the coding potential at the current position with the proper output value for the current position and uses this information to readjust its internal parameter weights. This continuous evaluation of the output error allows the network to optimize its performance.

The role of the neural network in this application is distinctly different from applications where nets have been trained to examine sequence data directly (15, 16). We have used the net to examine the sequence indirectly, through an array of information-rich sensors that make use of many different properties of the sequence. For example, for a neural network to learn the distribution of the 4096 6-mers in coding vs. noncoding DNA would require a very large network with hundreds of thousands of internal connections. Training such a network would be a long and difficult process. In practical terms, a nearly equivalent result could be obtained by maintaining the statistical preferences for 6-tuple words in a simple table and using a sensor algorithm to calculate the total preference over a window. The subsequent role of the neural net would be to combine this sensor with others, and not for the large-scale memorization of statistical facts. This simple concept makes practical the construction and training of recognition modules that incorporate large amounts and diverse types of information.

This coding module is not designed to find exons per se, but rather, exons, or portions of exons, that encode protein. Since the CRM is designed to find protein-coding regions, the 5' and 3' noncoding regions of exons are assigned as noncoding during the training process and are not expected to be found in the test sequences. The training set was constructed from 240,000 bases of sequence data. GenBank sequences were manipulated using the University of Wisconsin Genetics Computer Group software (17). Network training involved examination of 500,000 sensor output data vectors and the appropriate logical file representing the correct network outputs for each set of sensor data.

Evaluation of the CRM

The sensor algorithms in conjunction with the trained network form a single integrated unit designed to recognize coding sequence. Its performance was evaluated by examining 19 human test genes, none of which were included in the training set. Typical results of application of the CRM are shown in Figs. 2H and 3. The output of the CRM is characterized by its decisive nature and a very low level of noise. In our test set, the module located 90% (71/79) (Table 1) of coding exons of ≥100 bases. Only 20 of the 105 coding exons

---

†GenBank locus names of sequences used in neural network training were HUMACTGA, HUMADAG, HUMAFP, HUMALDBI, HUMALDCG, HUMBAR, HUMBSF2, HUMC056, HUMC060, HUMCRYGB, HUMAP0C2, HUMCVC1A, HUMERPA, HUMERPB, HUMERPC, HUMFESFPS, HUMFEBDA, HUMGASTA, HUMHCSA, and HUMGRP78; those used in testing were HUMALPHA, HUMAPRT, HUMPOS, HUMMETIA, HUMMCRA, HUMP425C17, HUMPAIA, HUMPLPSPC, HUMPEN1, HUMMTA, HUMMPORC, HUMPRCA, HUMPRPH1, HUMRASH, HUMSAA, HUMTBBS5, HUMTCRAC, HUMTHB, HUMTKRA, and HUMTPA.
calculate the sensors values, the trained network often does not sense a coding region until it is examining input from 2–3 positions (20–30 bases) inside a given coding exon, a point at which most of the window is within the coding region. Also, we might expect that the CRM would have difficulty correctly assigning coding positions in exons that are significantly shorter than 100 bases. In our test set, 30 of the coding exons (27.5%) were <100 bases long, and the module located 14 (47%) of these. This percentage of short exons is typical for the genome as a whole (18).

The performance of the CRM can also be evaluated in terms of the fraction of total positions in the sequence that are correctly assigned. Of the 17,576 sequence locations tested, 16,592 (96%) were correctly assigned as coming from either coding or noncoding DNA (Table 1). Of the 1113 test points classified as coding, 1029 (92%) were correctly assigned. The number of coding test positions that were missed because their coding exons (>100 bases long) were missed was only 87 (8%). The number of false positives (noncoding positions assigned as coding) was 84 (8%). Of the false positive assignments, 33 were contiguous with the 5’ or 3’ ends of actual coding regions, leaving the number of unexplained false positive assignments as 51 (5%). The 51 false positive positions were distributed among 20 false coding segments. Most of these false coding regions are short, containing <3 consecutive coding assignments. Any region indicated as coding in >3 consecutive positives is correctly assigned 94% of the time.

The use of neural networks to recognize features in DNA sequences has been pioneered by Lapedes and colleagues at Los Alamos National Laboratory. They have presented (16) a neural-network discriminator for distinguishing coding vs. noncoding sequences, which examined windows that were entirely coding or noncoding. Though this system performed the task quite well, a direct comparison of our results with theirs is difficult since they did not deal with windows that overlap edges of coding and noncoding regions. One striking difference between the results obtained by Lapedes et al. and our results is in the number of false positive indicated by the neural network. In one example in which Lapedes et al. (16) attempted to locate coding regions in an actual gene sequence, they found “many windows that are not known to be coding that the net indicates having coding character.” A CRM whose training set did not include this example sequence indicated only one false positive coding exon when examining this sequence. This low noise level is one of the key attributes of the multiple sensor–neural network approach. A low false positive rate is important for locating unknown genes in anonymous DNA sequences, since the analysis of a predicted coding region can be very time-consuming.

Our results thus far indicate little dependence on the precise network configuration, although performance is compromised with fewer nodes in the hidden layers. Other neural-network types and configurations, as well as additional sensors that consider sequence complexity and information content, need to be evaluated. With regard to the sensitivity of the CRM to errors in DNA sequencing, preliminary results suggest that the CRM can tolerate up to 1% insertion/deletion errors and still return useful information about the coding character of a region of sequence.

The CRM is being developed as a part of a larger artificial intelligence-based system designed to extract the maximum amount of information from anonymous DNA sequence. As a stand-alone module the CRM can predict the coding potential of a few hundred bases of DNA sequence with enough accuracy to justify using a portion of the sequence as a probe for gene expression, cDNA isolation, or comparing the translation of a predicted coding region with known protein sequences. Presumably, sequences with high coding poten-
Table 1. Performance of the CRM in locating coding regions in 19 human genes

<table>
<thead>
<tr>
<th>Locus</th>
<th>Total coding exons*</th>
<th>Coding exons indicated</th>
<th>Actual exons found*</th>
<th>Total test positions</th>
<th>No. of coding positions</th>
<th>Positions scored as coding</th>
<th>Positions correctly scored as coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMALPHA</td>
<td>11 (9)</td>
<td>11</td>
<td>11 (9)</td>
<td>446</td>
<td>163</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>HUMAPRT</td>
<td>5 (3)</td>
<td>6</td>
<td>5 (3)</td>
<td>292</td>
<td>55</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>HUMFOS</td>
<td>4 (4)</td>
<td>6</td>
<td>3 (3)</td>
<td>611</td>
<td>114</td>
<td>55</td>
<td>46</td>
</tr>
<tr>
<td>HUMMETIA</td>
<td>3 (0)</td>
<td>3</td>
<td>1 (0)</td>
<td>285</td>
<td>17</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>HUMNMYCA</td>
<td>3 (3)</td>
<td>3</td>
<td>3 (3)</td>
<td>867</td>
<td>164</td>
<td>100</td>
<td>94</td>
</tr>
<tr>
<td>HUMP45C17</td>
<td>8 (7)</td>
<td>7</td>
<td>7 (7)</td>
<td>845</td>
<td>157</td>
<td>92</td>
<td>89</td>
</tr>
<tr>
<td>HUMPAIA</td>
<td>8 (5)</td>
<td>8</td>
<td>6 (4)</td>
<td>1577</td>
<td>122</td>
<td>75</td>
<td>64</td>
</tr>
<tr>
<td>HUMPLPSPC</td>
<td>5 (1)</td>
<td>4</td>
<td>4 (3)</td>
<td>331</td>
<td>60</td>
<td>29</td>
<td>26</td>
</tr>
<tr>
<td>HUMPNNMTA</td>
<td>2 (2)</td>
<td>3</td>
<td>2 (2)</td>
<td>408</td>
<td>85</td>
<td>55</td>
<td>53</td>
</tr>
<tr>
<td>HUMPPOMC</td>
<td>2 (2)</td>
<td>2</td>
<td>2 (2)</td>
<td>856</td>
<td>80</td>
<td>72</td>
<td>70</td>
</tr>
<tr>
<td>HUMP RCA</td>
<td>7 (5)</td>
<td>6</td>
<td>5 (5)</td>
<td>1163</td>
<td>138</td>
<td>89</td>
<td>86</td>
</tr>
<tr>
<td>HUMPRPH1</td>
<td>3 (1)</td>
<td>2</td>
<td>1 (1)</td>
<td>485</td>
<td>50</td>
<td>33</td>
<td>29</td>
</tr>
<tr>
<td>HUMRASH</td>
<td>4 (4)</td>
<td>4</td>
<td>4 (4)</td>
<td>636</td>
<td>56</td>
<td>52</td>
<td>47</td>
</tr>
<tr>
<td>HUMSAA</td>
<td>3 (2)</td>
<td>2</td>
<td>2 (2)</td>
<td>336</td>
<td>37</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>HUMTBB5</td>
<td>4 (3)</td>
<td>5</td>
<td>4 (3)</td>
<td>878</td>
<td>133</td>
<td>127</td>
<td>118</td>
</tr>
<tr>
<td>HUMTCRAC</td>
<td>3 (2)</td>
<td>3</td>
<td>2 (2)</td>
<td>499</td>
<td>42</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>HUMTHB</td>
<td>14 (10)</td>
<td>11</td>
<td>10 (9)</td>
<td>2071</td>
<td>187</td>
<td>92</td>
<td>89</td>
</tr>
<tr>
<td>HUMTKRA</td>
<td>7 (4)</td>
<td>7</td>
<td>6 (3)</td>
<td>1340</td>
<td>70</td>
<td>37</td>
<td>34</td>
</tr>
<tr>
<td>HUMTPA</td>
<td>13 (10)</td>
<td>10</td>
<td>7 (6)</td>
<td>3650</td>
<td>169</td>
<td>56</td>
<td>49</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>109 (79)</strong></td>
<td><strong>105</strong></td>
<td><strong>85 (71)</strong></td>
<td><strong>17576</strong></td>
<td><strong>1895</strong></td>
<td><strong>1113</strong></td>
<td><strong>1029</strong></td>
</tr>
</tbody>
</table>

*In parentheses is the number of coding exons over 100 bases long.

We have used the CRM to examine DNA sequences generated from both human and mouse DNA in a number of laboratories. Thus far the system has been very successful in locating genes known to be within the sequenced regions and/or in predicting protein-coding regions that have been verified by the isolation of cDNA clones corresponding to the predicted region. The results of these studies will appear elsewhere.

---

We acknowledge the help and encouragement of other members of the computer sequence-analysis group at Oak Ridge National Laboratory, particularly Mr. Reinhold Mann, the project director, and Mr. Russell Hand, Jr., for compilation of 6-tuple statistics. This work was supported by the Office of Health and Environmental Research, U.S. Department of Energy, under Contract DE-AC05-84OR21400 with the Martin Marietta Energy Systems, Inc. We are particularly grateful for support from the Exploratory Studies Program of the Oak Ridge National Laboratory.