

FEB 12 1991

ORNL/TM-11741

ornl

ORNL
MASTER COPY

OAK RIDGE
NATIONAL
LABORATORY

MARTIN MARIETTA

**A Neural Network – Multiple Sensor
Based Method for Recognition of
Gene Coding Segments in Human
DNA Sequence Data**

E. C. Uberbacher
R. C. Mann
R. C. Hand, Jr.
R. J. Mural

MANAGED BY
MARTIN MARIETTA ENERGY SYSTEMS, INC.
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831; prices available from (615) 576-8401, FTS 620-8401.

Available to the public from the National Technical Information Service, U.S. Department of Commerce, 5085 Port Royal Rd., Springfield, VA 22161.

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/TM-11741

Engineering Physics and Mathematics Division

**A NEURAL NETWORK - MULTIPLE SENSOR BASED METHOD
FOR RECOGNITION OF GENE CODING SEGMENTS IN HUMAN
DNA SEQUENCE DATA**

E. C. Uberbacher*
R. C. Mann
R. C. Hand, Jr.*
R. J. Mural*

DATE PUBLISHED: February 1991

*Biology Division

Prepared by the
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831
managed by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-84OR21400

TABLE OF CONTENTS

	<u>PAGE NO.</u>
FIGURE LEGENDS	iv
ABSTRACT	v
INTRODUCTION	1
CODING RECOGNITION MODULE DESIGN OVERVIEW	2
SENSOR ALGORITHM DESCRIPTIONS	4
NEURAL NETWORK TRAINING AND CRM APPLICATION	6
CONCLUSION	7
REFERENCES	8

FIGURE LEGENDS

Figure 1. Schematic diagram showing the basic components of the coding recognition module. Seven sensor algorithms are used to characterize different attributes of the DNA sequence related to coding probability within a 100 base pair sequence window. These signals are input to a neural network trained to interpret them and make a decision regarding the presence of coding DNA. The entire structure represents one recognition module of the GENESIS system.

Figure 2. Signals produced by the seven sensors described in the text for the 6500 base pair human ras proto-oncogene sequence region. The signal amplitudes at each given sequence position are used by the neural net to predict the coding probability corresponding to that position.

Figure 3. Examples of output predictions for several regions of test DNA. The GENE BANK sequence entry names are indicated, and the arrows correspond to the actual coding segment positions in these sequences.

ABSTRACT

A central focus of the Human Genome Project is to locate and characterize genes within genomic DNA sequence data. Of particular importance are genes related to cancer and other major human genetic diseases. Identification of gene segments within the DNA sequence currently relies upon tedious and time-consuming experimental methods. A new computational method developed at Oak Ridge National Laboratory [ORNL] represents an efficient alternative to these experimental procedures. This method combines a set of statistical sensors and a neural network into a single structure which is capable of locating gene segments with considerable speed and accuracy. The approach provides a powerful tool for investigators searching for disease genes.

INTRODUCTION

One of the major challenges facing the Human Genome Project is to develop the computing technology necessary to analyze vast amounts of raw DNA sequence data for biologically important features. Major sequencing efforts, using currently available technology, are already underway in several laboratories, and are producing significant amounts of DNA sequence data for both the human genome and other model organisms. Several of these programs are directed toward localization and identification of major human disease genes. A fundamental problem in these studies is locating within the sequence the segments constituting the genes. Even very precise genetic mapping, using current methodology, can only localize a disease gene to a region of several million base pairs. Since the gene product of most disease genes is not known, cDNAs cannot be used to gain access to the exact chromosomal location of the genes of interest. As a result, current experimental protocols for locating coding regions are highly labor intensive and time consuming. Furthermore, currently available computational methods for recognition of coding DNA sequence regions have proved to be unreliable.

The purpose of this report is to describe a new computational methodology developed at ORNL, which locates coding segments accurately and efficiently. This methodology can be of significant benefit to many investigators currently analyzing sequence data in search of genes, including those related to cancer and other genetic diseases.

The approach described in this report is being used to solve a wide range of sequence feature recognition problems within the framework of an ongoing ORNL research and development project. The overall goal of this project is to combine parallel computing, artificial intelligence, and molecular biology based technologies to produce a high-speed integrated artificially intelligent system for the location, identification, and interpretation of genes, regulatory regions, and other biologically important features in Human Genome sequence data. This system, called the Genetic Sequence Interpretation System, or GENESIS, will take several years to construct. In the interim, however, certain modular parts, which can function in a stand-alone manner, can be made available as tools to aid current sequence analysis efforts. The "coding recognition module", or CRM, described in this report, is the modular block of GENESIS designed to locate protein coding segments within sequence data, and demonstrates the principles being used to recognize other sequence features as well.

CODING RECOGNITION MODULE DESIGN OVERVIEW

We have approached the problem of DNA sequence data interpretation as a pattern recognition process utilizing information from multiple sensors. This is analogous to the situation in sensor-based robotic systems where the perception of the robot's surroundings occurs through integration of information from multiple sensors, e.g. CCD cameras, sonar transducers, laser range finders, tactile sensors, etc. These sensors supply partially redundant information with different levels of accuracy and uncertainty depending on the state of the environment being sensed. By optimally integrating the information presented by the sensors, a combined best estimate of the environment can be obtained that is better than estimates based on individual sensors alone.

This basic scheme, translated into the realm of DNA sequence analysis, provides the basis for the recognition module in GENESIS. The CRM described in this report, for example, incorporates a group of seven sensor algorithms (described below), each designed to provide an indication of a sequence property related to coding probability. Integration of the sensor algorithm outputs is accomplished with a neural network, which has been subjected to a training procedure to learn how to interpret this information, and subsequently make a decision about the presence of coding DNA. The idea is to use the neural network's learning ability to develop the most reliable indicator possible from a complex and rich array of sensory input. The CRM, and other sequence feature recognition modules, will each have independent learning capabilities, and each undergo its own training process to optimize its function. Figure 1 shows a schematic diagram of the CRM.

There is an important conceptual difference between the way neural networks are used here for sensor integration, and the approach used by several other investigators using neural networks for DNA sequence analysis (1). Typically, neural networks are trained to recognize DNA sequence patterns by direct examination of the DNA sequences corresponding to the desired features. The net is expected to learn what is necessary for feature recognition using only the base pairs (letters) in the example sequences used in training. However, there are limitations on what nets can extract in this manner. Some improvement in learning from direct sequence examination may be possible with very large networks, though such networks are likely to be unmanageable and learn very slowly.

The alternative demonstrated by the CRM is to show examples of the DNA sequence data to the neural network indirectly, filtered through a series of sensors designed to reveal the important characteristics of the sequence data to the net. The addition of this intervening layer of sensors gives the net a leg up on the problem, strengthening the network's recognition capabilities and increasing its learning speed. This approach to DNA sequence pattern recognition is unique to the ORNL effort, and is likely to have wide applicability to DNA sequence feature recognition.

The basic module design contains considerable flexibility in that additional or alternative sensors can be incorporated into a module with relatively simple alterations to the neural

net, and a subsequent retraining process. The structure of the overall GENESIS system is not affected. The design therefore contains the ability to incorporate parallel developments in sensors in a number of different laboratories, and unify them within a single framework.

SENSOR ALGORITHM DESCRIPTIONS

The recognition module designed to identify coding segments in Human genomic DNA (shown in Figure 1) currently consists of seven semi-independent algorithms designed to provide indicators related to the presence or absence of a coding region, and a neural network which has been trained to interpret the sensor algorithm outputs. The sensor algorithms evaluate a number of different attributes of coding DNA including triplet positional biases resulting from codon usage, sequence fractal characteristics, k-tuple vocabularies, and the presence and length of open reading frames.

An important consideration is that the characteristics of k-tuple usage, reading frame bias, and other attributes of DNA are organism specific. As a result, recognition processes may be strengthened by specializing the construction of modules, such as the CRM, for a specific DNA type; in this case, human DNA. It is a relatively simple task to create similar modules for other species, simply by incorporating the appropriate statistical information for the organism's k-tuple usage, bias, etc., into the standard sensor algorithms, and training the neural net.

To evaluate the probability that a given sequence position is within a coding segment, the seven algorithms are evaluated in a 100 base pair sequence window centered at the position, and the sensor signals are then evaluated by the neural net. A brief overview of the sensor algorithms follows:

- (1) Frame Bias Matrix: This method seeks to identify coding regions based on the non-random frequency with which each of the four bases occupies each of the three positions within codons, due to unequal use of amino acids, and preferred use of codons related to such factors as overall DNA composition (2). This bias, expressed as a matrix, is used as a probe to identify potential coding regions and preferred reading frame from the correlation coefficient between the matrix and each 100 base DNA window.
- (2) Fickett: This algorithm is a modification of an algorithm developed by Fickett (3), and is a measure of total triplet positional bias. The algorithm window as used here is changed from 200 base pairs to 100 base pairs.
- (3) Open Reading Frame: At each sequence position, the output value of this algorithm is proportional to the length of the open reading frame for the preferred frame as predicted by the Frame Bias Matrix algorithm above.
- (4) Dinucleotide Fractal Dimension: Dinucleotide occurrence is known to be far from random, with dinucleotides such as AA, TC, being common and CG being rare. It is thus possible to view a DNA sequence as a dynamic function by considering changes in energy in the Boltzmann sense, and using the energy scale $E = -\ln(p)$, where p is each dinucleotide's probability. These fluctuations can be characterized by a fractal dimension (2,4). Coding DNA usually has lower dimension than non-coding.

(5) Coding Segment 6-tuple word preferences: One way of characterizing the sequences of the human genome is to assemble an annotated compilation of all the k-tuple "words" of a given length, noting their frequency of occurrence in various feature types. A "sequence dictionary" has been constructed at ORNL which contains statistics for the usage of 4-, 5-, 6-, 8-, and 10-tuple words in coding segments compared to non-coding DNA, and also in other contexts. In this sensor algorithm, within a 100 base pair sliding window, the observed 6-tuple word preferences are scored and summed to provide a coding indicator. Each word is scored by the log ratio of its occurrence in coding vs. non-coding DNA. An additional sensor algorithm which examines the preferences for 6-tuple words occurring in the appropriate reading frame is being constructed and will represent an eighth sensor for the module.

(6) Intron 5-tuple word preferences: Similar to the preceding algorithm except statistics from the sequence dictionary for intronic DNA vs. bulk DNA are used. This output represents a negative coding indicator. An additional sensor is being designed which contrasts intron vocabulary with coding vocabulary, and this should contribute to proper coding segment edge detection.

(7) Repetitive 5-tuple word preferences: Similar to preceding two algorithms except that 5-tuple dictionary statistics for various classes of repetitive DNAs are used in comparison to coding DNA. This is also a negative coding indicator.

As an example of the application of these seven algorithms, Figure 2 shows the sensor outputs for the 6500 base pair human ras proto-oncogene region.

NEURAL NETWORK TRAINING AND CRM APPLICATION

A backpropagation neural network (5), consisting of 7 input nodes, two hidden layers of 12 and 5 nodes, and an output node, was used to integrate the sensor data. In the training procedure, the seven sensor algorithms were applied to 240 kilobases of human genomic DNA sequence data containing 24 genes and 149 exons, along with intronic and flanking DNA, etc. Input training vectors were calculated at consecutive positions every 10 base pairs along the sequence, and the sensor algorithm outputs supplied to the neural net, along with a logical input (1 or 0) to indicate whether the input vector was generated from a coding or non-coding segment. The net was subjected to 1×10^6 training examples.

To test learning, the entire CRM (sensors plus net) was applied to a number of DNA sequence regions not used in the training set. Typical results are shown in Figure 3. The first test output corresponds to the sensor inputs shown for the human ras proto-oncogeny region in Figure 2. Despite the apparent complexity of the sensor data, the CRM output is very decisive and very noise free. In the module output for the five test genes, virtually all significant peaks correspond to actual coding segments (shown by arrows). In most cases, the extent of the coding segments in the sequence is also well predicted. In its present form, the CRM locates more than 90% of all coding segments in the tested human DNA, including a significant percentage of coding segments shorter than 100 base pairs. Furthermore, even with application of a fairly low significance threshold, less than 10% of the observed peaks correspond to non-coding regions. With the addition of the two sensors under construction, and other minor refinements, it is expected that the sequence window size used by the sensors can be reduced from 100 base pairs to 50 base pairs, making recognition of very short coding segments (<50 base pairs) quite reliable. Further experiments are also planned with other neural net types and configurations.

CONCLUSION

The foundation of the sequence analysis process is necessarily built on its ability to recognize feature related patterns in DNA sequence data. Examining DNA sequence regions with sensor algorithms, and then integrating the outputs of these algorithms with neural networks, has proved to be a powerful mechanism for strengthening the recognition process. We have demonstrated that neural networks are capable of combining the outputs of coding DNA sensor algorithms to provide a combined indicator that is better than 90% correct for human genomic DNA. This same technology is being applied at ORNL to many other DNA sequence pattern recognition problems. In its present form, the CRM represents a powerful tool which could greatly aid current experimental efforts to locate important human genes.

REFERENCES

1. Brunak, S., Engelbrecht, J., and Knudsen, S., Nucl. Acids Res., 18, pp. 4797-4801 (1990).
2. Mural, R. J., Mann, R. C., and Uberbacher, E. C., Proceedings of the First International Conference on Electrophoresis, Supercomputing, and the Human Genome (1990).
3. Fickett, J., Nucl. Acids Res., 10, pp. 5303-5318 (1982).
4. Hsu, K., and Hsu, A., Proc. Natl. Acad. Sci., USA 87, pp. 938-941 (1990).
5. Hopfield, J. (1982) P.N.A.S., 79, pp. 2554-2558 (1982); Hopfield, J., P.N.A.S., 84, pp. 8429-8433 (1987).

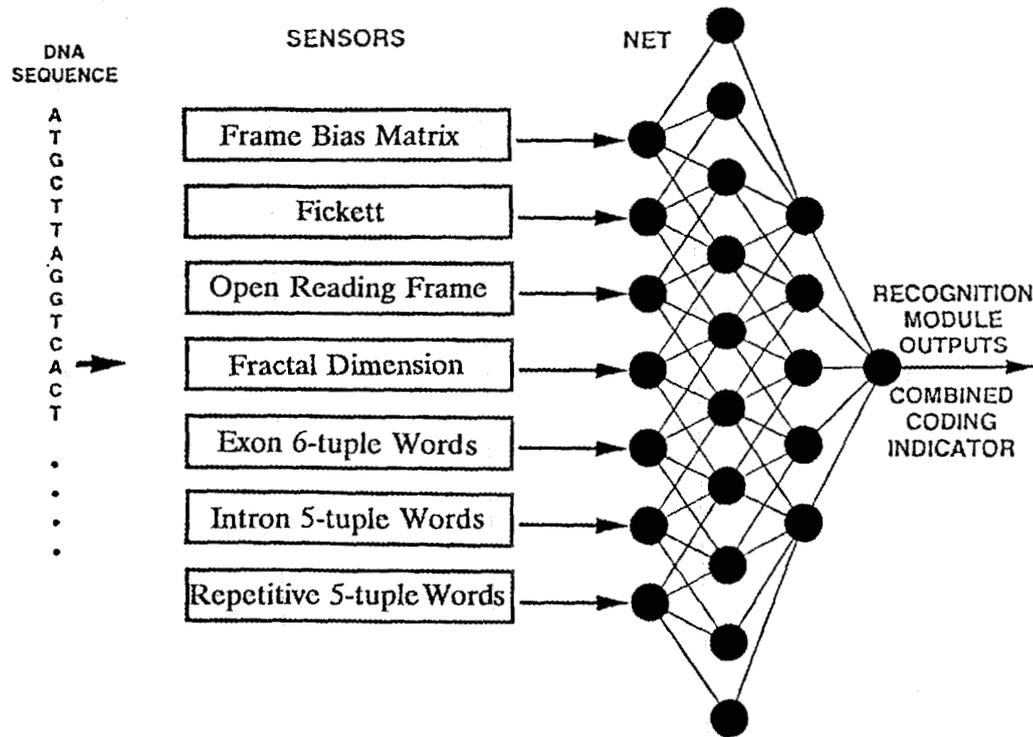


Figure 1. Schematic diagram showing the basic components of the coding recognition module. Seven sensor algorithms are used to characterize different attributes of the DNA sequence related to coding probability within a 100 base pair sequence window. These signals are input to a neural network trained to interpret them and make a decision regarding the presence of coding DNA. The entire structure represents one recognition module of the GENESIS system.

SENSORS

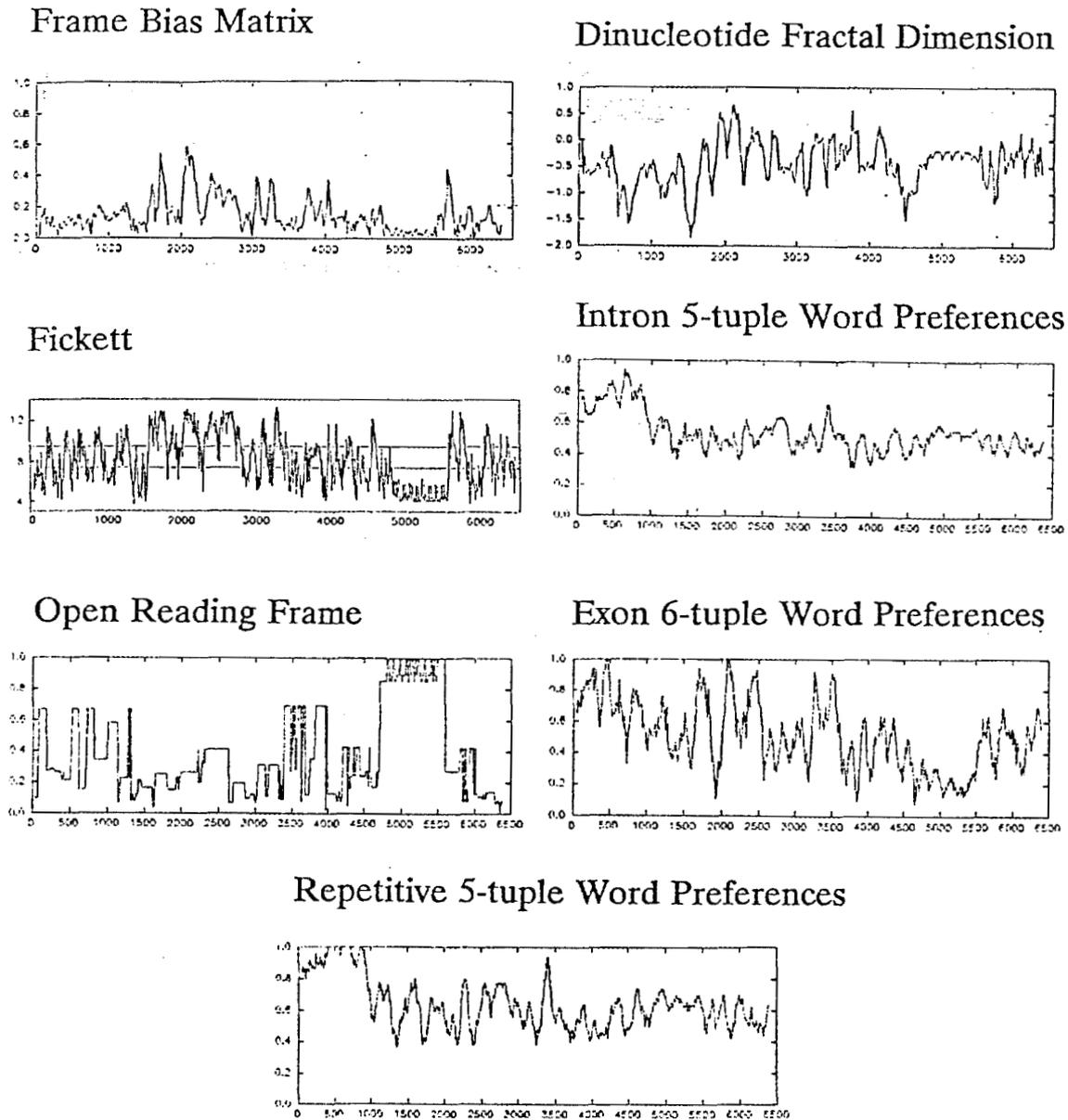


Figure 2. Signals produced by the seven sensors described in the text for the 6500 base pair human ras proto-oncogene sequence region. The signal amplitudes at each given sequence position are used by the neural net to predict the coding probability corresponding to that position.

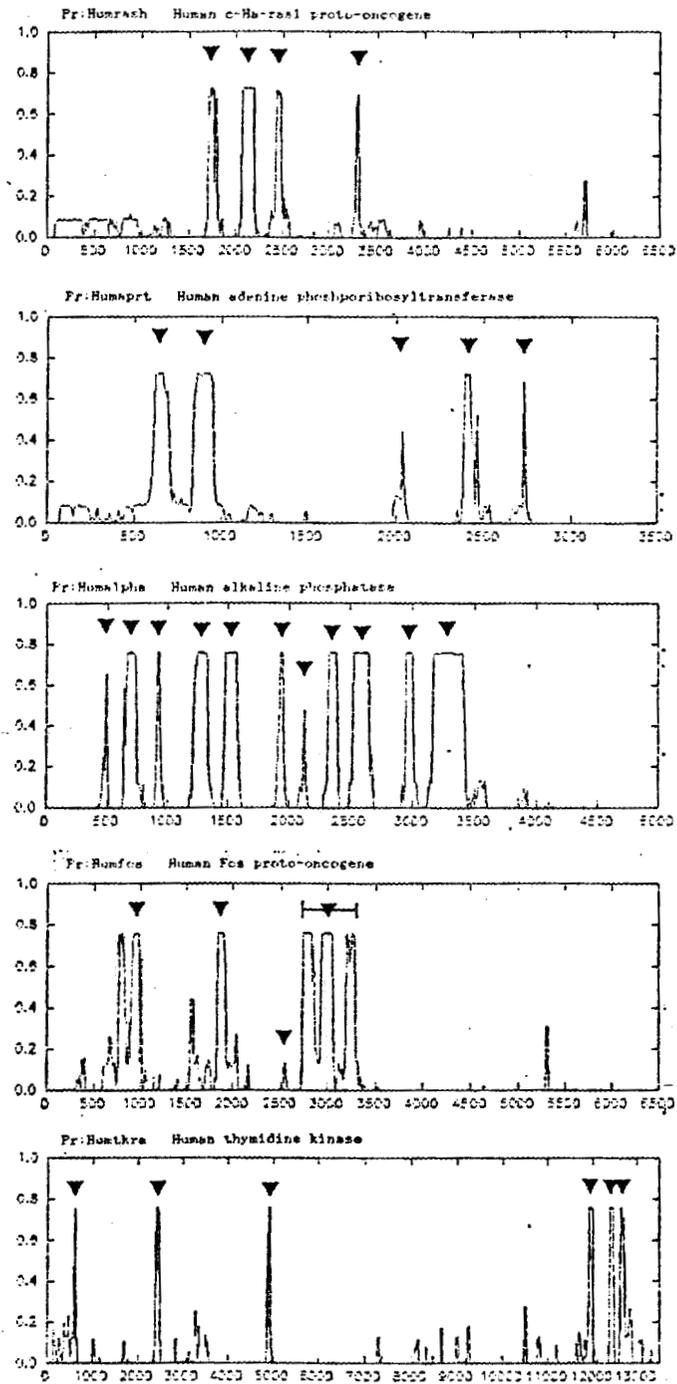


Figure 3. Examples of output predictions for several regions of test DNA. The GENE BANK sequence entry names are indicated, and the arrows correspond to the actual coding segment positions in these sequences.

THIS PAGE LEFT BLANK INTENTIONALLY.

INTERNAL DISTRIBUTION

- | | | | |
|--------|-------------------|--------|-------------------------------|
| 1. | B. R. Appleton | 30. | R. C. Ward |
| 2. | J. R. Einstein | 31. | J. J. Dorning (Consultant) |
| 3. | X. Guan | 32. | J. E. Leiss (Consultant) |
| 4- 8. | R. C. Hand, Jr. | 33. | N. Moray (Consultant) |
| 9. | F. C. Hartman | 34. | M. F. Wheeler (Consultant) |
| 10. | J. P. Jones | 35. | EPMD Reports Office |
| 11. | F. C. Maienschein | 36-37. | Laboratory Records Department |
| 12-16. | R. C. Mann | 38. | Laboratory Records, ORNL RC |
| 17-21. | R. J. Mural | 39. | Document Reference Section |
| 22. | C. E. Oliver | 40. | Central Research Library |
| 23. | D. E. Reichle | 41. | ORNL Patent Office |
| 24. | C. R. Richmond | | |
| 25-29. | E. C. Uberbacher | | |

EXTERNAL DISTRIBUTION

42. Office of the Assistant Manager for Energy Research and Development, Oak Ridge Operations, U.S. Department of Energy, P.O. Box 2008, Oak Ridge, TN 37831.
43. B. J. Barnhart, Program Manager, Human Genome Program, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, Washington, DC 20585.
44. J. Schmaltz, Human Genome Program, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, Washington, DC 20585.
45. G. Featheringham, Cray Research, Inc., Suite 1331N, 1331 Pennsylvania Avenue, NW, Washington, DC 20004.
46. J. Hache, Scientific Director, Bertin et Compagnie, 59, rue Pierre Curie, Z.I. des Gatines B.P. No.3, 78373 Plaisir Cedex, France.
47. T. Sumitani, Systems Engineering Division, Hitachi, Ltd., 6, Kanda-Surugadai, 4-Chome, Chiyoda-Ku, Tokyo, 101 Japan.
48. M. S. Hutchinson, Computation Department, Lawrence Berkeley Laboratory, Berkeley, CA 94720.
49. M. D. Zorn, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720.
50. F. Olken, Computer Science Research Department, Lawrence Berkeley Laboratory, Berkeley, CA 94720.

ORNL/TM-11741

51. J. Fickett, Los Alamos National Laboratory, Los Alamos, NM 87545.
52. A. Loh, Scientific Applications Marketing Group, Digital Equipment Corporation, Marlboro, MA 01752-9102.
53. C. Watanabe, Genetech, Inc., 1861 Drake Drive, Oakland, CA 94611.
54. R. Balderalli, Department of Genetics, Harvard Medical School, Howard Hughes Medical Institute, Boston, MA 02115.
55. M. J. Gliwias, 2180 Radcliffe Drive, Westlake, OH 44145.
56. H. Delius, DKFZ ATV, I.N.F. 506, 69 Heidelberg, Republic of Germany.
57. S. Stefan, Biology Department, Indiana University, Bloomington, IN 47405.
- 58-67. Office of Scientific and Technical Information, P. O. Box 62, Oak Ridge, TN 37831.