MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0358511 8

ORNL/TM-11924

ornl

OAK RIDGE
NATIONAL
LABORATORY

MARTIN MARIETTA

GAP—A Computer Program for Gene Assembly

J. R. Einstein
E. C. Uberbacher
X. Guan
R. J. Mural
R. C. Mann

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
CIRCULATION SECTION
4500N ROOM 175
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this report, send in name with report and the library will arrange a loan.

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 626-8401.

Available to the public from the National Technical Information Service, U.S. Department of Commerce, 5285 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-11924

Engineering Physics and Mathematics Division

GAP - A COMPUTER PROGRAM FOR GENE ASSEMBLY

J. R. Einstein, E. C. Uberbacher,* X. Guan, R. J. Mural,* and R. C. Mann

*Biology Division.

DATE PUBLISHED - September 1991

**Research supported by the Office of Health and Environmental Research,
U.S. Department of Energy, and the
Laboratory Directed Research and Development Programs**

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

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0358511 8

CONTENTS

	Page No.
ABSTRACT	v
1. INTRODUCTION	1
2. METHODS	2
3. CONSIDERATION OF ERRORS	4
4. TESTS OF GAP	5
5. CONCLUSIONS	7
ACKNOWLEDGMENTS	8
REFERENCES	9

ABSTRACT

A computer program, GAP (Gene Assembly Program), has been written to assemble and score hypothetical genes, given a DNA sequence containing the gene, and the outputs of several other programs which analyze the sequence. These programs include the coding-recognition and splice-junction-recognition modules developed in this laboratory. GAP is a prototype of a planned system in which it will be integrated with an expert system and rule base. Initial tests of GAP have been carried out with four sequences, the exons of which have been determined by biochemical methods. The highest-scoring hypothetical genes for each of the four sequences had percent correct splice junctions ranging from 50 to 100% (average 81%) and percent correct bases ranging from 92 to 100% (average 96%).

1. INTRODUCTION

A major challenge facing the Human Genome Project is to develop the technology to recognize and interpret the information contained in genomic sequence data. Although large-scale sequencing efforts are just beginning, a number of laboratories are already producing 100s of kilobases of DNA sequence, and would greatly benefit from reliable computational methods to locate exons and other gene components. The difficulty in localizing genes within large blocks of sequence data is a major impediment to this community and is discouraging sequencing efforts (1). Experimental methods used to localize genes, such as exon trapping, zoo-blots, and recombination methods, are labor intensive and subject to interpretation problems.

Although recognizing features from DNA sequence data is difficult, we know that the problem is theoretically tractable since organisms utilize the signals found in DNA with great accuracy. Certain types of information - for example, the 64 triplets which constitute the genetic code - are well understood. Unfortunately, our understanding of many other types of information is not so complete. The use of consensus sequences or statistical matrices for identifying certain features of interest is often complicated by the problem of large numbers of false signals which occur by chance (2). For example, in the 6.5 kb c-Ha-ras1 proto-oncogene entry in GENBANK, there are several hundred occurrences of the triplet YAG (the three bases conserved at an acceptor splice junction), yet only three of these represent actual splice junctions.

As a result of difficulties with reliable feature recognition, few investigators have even attempted the construction of computer assisted, let alone fully automated, systems for the assembly of gene components into high-level objects such as genes. Perhaps the most notable of these is GM (3), which compensates for noise in the recognition process by applying biologically necessary constraints to the assembly process. GM's primary mechanism for recognizing exons, the identification of open reading frames, is of limited use for eukaryotic DNA. Despite this and other limitations, GM represents an important step, primarily because of the integration of multiple constraints in the assembly process.

The use of artificial-intelligence and machine-learning techniques to improve pattern recognition with DNA sequences is a relatively recent development. Direct interpretation of sequence data by neural networks has been used in several pioneering studies on coding DNA and splice junction recognition at LANL (4). A group in Sweden has also recently developed a neural-net-based system for splice junction recognition (5) and for recognition of coding regions (6). We have taken a somewhat different approach, in that the sequence is first processed with a group of mathematical algorithms or "sensors", and the neural network learns to recognize features from the output of these sensors, rather than by examining the sequence directly (7). This approach has certain advantages in that smaller networks are required, learning is faster, and a variety of types of information can be combined in the analysis.

The research to be described utilizes the results of the recognition tools developed for coding exons and splice junctions as the basis for a prototype system for automated gene assembly. This C-language prototype system is designed to be used later in conjunction with an expert system. The purposes of this prototype are to test the effectiveness of the present feature-recognition tools, to develop principles for automated gene assembly, and to explore error sources and correction strategies.

2. METHODS

The computer program GAP (Gene Assembly Program) identifies and assembles the exons of a gene on the basis of the following input:

a) a base sequence, known or assumed to contain one gene;

b) a coding-probability (CP) function over the sequence;

c) a probable-reading-frame (PRF) function over the sequence;

d) the positions and scores of possible donor and acceptor splice junctions in the given sequence;

e) the positions of possible translation-initiation sites.

The source of inputs (b) and (c) is the coding-recognition module (CRM) under development in this laboratory (7). CRM now incorporates a group of seven "sensor" algorithms, each designed to provide an indication of the coding potential of a given region of sequence, and a neural network with one input node for each sensor and one output node. Through training, the neural network adjusts its internal weights so as to recognize correlations between sensor outputs which correspond to sequence properties of coding DNA.

The sources of inputs (d) are acceptor and donor splice-junction-recognition modules also under development in this laboratory (8). Inputs (e) were obtained with the use of an algorithm similar to that discussed by Kozak (9).

The procedure with which GAP assembles possible genes comprises the following steps:

(1) The approximate edges, or limits, of coding regions are obtained by inspection of the CP function.

(2) An attempt is made to reduce the set of possible reading frames (defined relative to the unspliced sequence) for each coding region (CR), on the basis of the PRF function and the presence or absence of stop codons within the "heart" of the

CR, for each possible reading frame.

(3) Sets of possible donors and acceptors, taken from input (d) above, are assigned to each CR, based on their positions relative to the edges of the CR.

(4) All possible genes consistent with the input information are assembled and scored.

Initially, each possible gene consists of the CP-function-based set of CR's, spliced by a combination of the possible donors and acceptors, subject to following hypotheses:

(a) no stop codons exist in the assembled gene;
(b) the reading frame (relative to the unspliced sequence) of each CR agrees adequately with the PRF function;
(c) exons have a certain minimum length (a parameter of the program);
(d) introns have a certain minimum length (a parameter).

Furthermore, for a hypothetical gene to be accepted as possible, it must have a translation-initiation site, taken from input (e), the position of which satisfies two conditions: being within an "acceptable" distance from the left edge of the leftmost CR, and agreeing with the reading frame assigned to that CR during the splicing procedure.

The end of the hypothetical gene is the first stop codon encountered beyond the left edge of the rightmost CR, given its reading frame.

In this initial version of GAP, the score of a hypothetical gene is taken to be simply the geometric mean of the scores of all utilized splice junctions.[1]

A general problem of assembling genes in the manner outlined above is the large number of possible combinations of splice junctions to be considered. An example is the recently investigated sequence of HUMALPHA (Genbank: Human Alkaline Phosphatase), containing eleven exons, for which GAP found about 5×10^{10} combinations of possible splice junctions. GAP greatly reduces the computing burden by (1) splicing in one direction at a time from the best-determined CR, for one of its possible reading frames; (2) discontinuing splicing beyond any junction which implies a violation of one of the above hypotheses; (3) keeping a record of splices which have been ruled out, in case the splicing process returns by a different route to junctions considered previously.

---

[1]For this purpose, the input score of a splice junction is multiplied by a factor dependent on the distance of the junction from its statistically expected position relative to the initially determined edge of the CR. That is, a splice junction which has a high score because of close coincidence with its consensus sequence receives a lower composite score if it is far outside the edge of the CR or deeply within it.

The following two examples of the methods by which GAP "prunes" the search tree pertain to splicing to the right. A donor junction positioned to the right of a CR is not accepted if there is a stop codon (in the reading frame already assumed for that CR) between the CR and the junction. Then, an acceptor junction for the next CR to the right is not accepted if, given the already assumed donor, an "impossible" reading frame is implied for the next CR.

By these methods, and because the number of possible reading frames had been reduced to one for all the CR's, the number of individual splice-junction tests for HUMALPHA was reduced to about 1.3×10^6, which were done on an IBM PC/AT in less than twelve minutes.

3. CONSIDERATION OF ERRORS

Sometimes, the initial version of GAP finds the correct gene, i.e., the set of exons determined biochemically, as the highest-scoring hypothetical gene [see results for HUMRASH (Genbank: C-Ha-Ras1 Proto-Oncogene)]. However, the highest-scoring hypothetical gene - or even all the hypothetical genes produced by GAP - may disagree in part with the biochemical results. This can occur because of imperfections in the methods used presently to provide GAP's input. Although there is a continuous effort to improve these methods, they will never be perfect. Therefore, it is desirable that the gene assembler include methods to detect errors and deal with them.

An initial step toward dealing with errors has been to include in GAP procedures for the insertion and deletion of CR's, and for testing all combinations of a set of proposed insertions and deletions. So far, criteria for the automatic proposal of insertions and/or deletions have not been sufficiently developed for use, so that the user must include them in the program input (but not specification of the splice junctions for inserted CR's).[2] It is planned that in future versions of GAP, CR changes for testing be proposed automatically on the basis of (a) the results with the original set of CR's; (b) previously unexamined features of the input data.

One case which GAP already handles automatically is that which arises when the program finds no satisfactory translation-start site for any "possible" reading frame of the leftmost CR. Then the possibilities of a CR insertion before the original leftmost CR, and/or

---

[2]It would not have been difficult to include code within GAP such that the desired CR insertion and CR deletion for HUMTKRA, and the desired CR deletion for HUMAPRT (see Table 1) would have been "automatic." The data and results for these two sequences contain features, indicating the possibility of errors, on the basis of which such code could have been written. However, we realize that experience with many sequences will be required before we can write code, and/or an expert-system rule base, adequate for proposing CR changes for the general case. At this point, it is more appropriate to make the desired changes part of the input to GAP.

deletion of that CR, are tested.

The insertion/deletion methods have so far succeeded completely in one of the three cases tested [HUMTKRA (Genbank: Human Thymidine Kinase)]. However, in two other cases [HUMALPHA and HUMAPRT (Genbank: Human Adenine Phosphoribosyltransferase)] the "correct" set of exons was not obtained. The reason in each case was that one of the biochemically determined splice junctions was not in GAP's list of possible junctions. In one case, the input score was too low. In the other, the algorithm of the program which prepares the donor list [input (d)] did not identify the sequence at the site in question as consistent with that of a donor splice junction. (Clearly, there may be cases, now and in the future, in which the error lies in the biochemical determination.)

We anticipate that very complex methods will be required for dealing adequately with errors in the gene-assembly process. For this purpose we are preparing to integrate future versions of GAP with an expert system, which should facilitate complex reasoning and the adoption of different strategies for different, unpredictable circumstances.

While it is necessary to try to eliminate errors, it should be noted that somewhat imperfect gene determinations are very useful to biochemists.

4. TESTS OF GAP

Table 1 shows the results of tests of GAP with four sequences. In all cases GAP assembled most of the gene correctly. In one case (HUMRASH) the highest-scoring gene agreed exactly with that determined biochemically - i.e., all the exons were found, and the positions of the translation-initiation site, splice junctions, and end of the last exon were the same. In another case (HUMTKRA) complete agreement was obtained after user specification of two CR changes: one insertion (splice junctions not specified) and one deletion.

In two cases (HUMALPHA and HUMAPRT) GAP did not find the completely correct gene with any score, even after CR changes for HUMAPRT. Tests were made with the input data slightly altered so as to make it possible for GAP to find the completely correct gene. The aim was to see whether the correct gene would be found, and how high it would score. Table 1 includes descriptions of the alterations.

TABLE 1. TESTS OF GAP

Name	HUMALPHA	HUMAPRT	HUMRASH	HUMTKRA
Sequence length	4556	3016	6453	13500
No. of exons	11	5	4	7
Highest-scoring gene (unmodified input)[3,4]				
No. exons missed	1	0	0	0
No. false exons	1	1	0	0
No. splice-junction differences[5]	1	4	0	0
Percent bases correct	94	92	100	100
Percent bases false	4	34	0	0
Percent splice junctions correct	75	50	100	100
Highest-scoring gene (modified input)[6]				
No. exons missed	0	0		
No. false exons	0	0		
No. splice-junction differences	0	3		
Place among highest scoring genes of "correct" gene	1	12		
Percent bases correct	100	86		
Percent bases false	0	3		
Percent splice junctions correct	100	63		
No. combinations of splice junctions	5.1×10^{10}	3.5×10^{4}	192	2.3×10^{4}
No. splice-junction tests	1.3×10^{6}	1150	93	862

[3]Proper scoring of hypothetical genes which have inserted and/or deleted CR's is still in a rudimentary state of analysis.

[4]For HUMAPRT the user specified the testing of one CR deletion (see discussion above); the highest-scoring gene was obtained with that deletion. For HUMTKRA the user specified one CR deletion and one CR insertion. The program automatically tested these changes and their combination, and it was with the combination that the highest-scoring hypothetical gene was obtained.

[5]For exons which match real exons.

[6]Modification to input for HUMALPHA: score of one input acceptor raised from 0.005 to 0.1 (maximum score: 1.0). Modification to input for HUMAPRT: entered an input donor splice junction with score 0.5, and increased score of an input acceptor splice junction from 0.01 to 0.9.

5. CONCLUSIONS

The computer program GAP assembles and scores hypothetical genes on the basis of a DNA sequence and the output of other programs developed in this laboratory to identify coding regions, probable reading frames, the positions of splice junctions, and the positions of translation-initiation sites. GAP identifies the approximate limits of coding regions and tests combinations of possible splice junctions in order to specify exactly the limits of each exon. A score is provided for each hypothetical gene.

GAP has been tested on four sequences, yielding in each case most or all of the correct gene. For the HUMRASH sequence, the highest-scoring hypothetical gene agreed exactly with the set of exons determined biochemically.

For another sequence (HUMTKRA) exact agreement was found after intervention by the user to direct the program to test the deletion of one coding region (CR) and the insertion of another (splice junctions not specified). The program automatically tested these two changes individually and in combination. The highest-scoring gene, obtained with the combination of the two changes, agreed completely with the biochemically determined set of exons. It is planned that in future versions of GAP these CR changes be proposed automatically.

For two additional sequences (HUMALPHA and HUMAPRT), the highest-scoring gene had a few errors, even after, in the case of HUMAPRT, specification of a CR deletion. Tests were made with slightly altered input data in order to investigate how GAP would perform. In both cases GAP found the completely correct gene; in one case, as the highest-scoring hypothetical gene. We intend to continue to improve our methods for identifying coding regions and splice junctions.

We plan to incorporate much more complex reasoning about possible errors by combining future versions of the program with an expert system.

ACKNOWLEDGMENTS

This research was supported by the Office of Health and Environmental Research, U. S. Department of Energy under Contract No. DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. The authors would like to particularly acknowledge support from the Laboratory Directed Research and Development Programs.

REFERENCES

1. *Science* 252, pp. 1618-1619 (1991).

2. R. Staden, *Nucl. Acids Res.* 12, pp. 505-519 (1984).

3. D. A. Fields and C. A. Soderlund, *CABIOS* 6, pp. 263-270 (1990).

4. A. Lapedes, C. Barnes, C. Burks, R. Farber and K. Sirotkin, in "Computers and DNA," *SFI Studies in the Sciences of Complexity*, G. Bell and T. Marr, eds., Vol. VII, pp. 157-182, Addison-Wesley (1989).

5. S. Brunak, J. Engelbrecht and S. Knudsen, *Nucl. Acids. Res.* 18, pp. 4797-4801 (1990).

6. S. Brunak, J. Engelbrecht and S. Knudsen, *J. Mol. Biol.* 220, pp. 49-65 (1991).

7. E. C. Uberbacher and R. J. Mural, *Proc. Natl. Acad. Sci.* USA (in press).

8. R. J. Mural, R. C. Mann, and E. C. Uberbacher, in *Proc. First International Conference on Electrophoresis, Supercomputing and the Human Genome*, C. R. Cantor and H. A. Lim, eds., pp. 164-172, World Scientific Publishing Co. (1991).

9. M. Kozak, *J. Cell Biol.* 108, pp. 229-241 (1989).

INTERNAL DISTRIBUTION

1. B. R. Appleton
2- 6. J. R. Einstein
7. N. Wright Grady
8-11. X. Guan
12. R. E. Hand, Jr.
13. F. C. Hartman
14. J. P. Jones
15-19. R. C. Mann
20. T. J. Mitchell
21-25. R. J. Mural
26. C. E. Oliver
27. D. E. Reichle
28. C. R. Richmond
29-33. E. C. Uberbacher
34. R. C. Ward
35. R. W. Brockett (Consultant)
36. J. J. Dorning (Consultant)
37. J. E. Leiss (Consultant)
38. N. Moray (Consultant)
39. M. F. Wheeler (Consultant)
40. EPMD Reports Office
41-42. Laboratory Records Department
43. Laboratory Records, ORNL RC
44. Document Reference Section
45. Central Research Library
46. ORNL Patent Office

EXTERNAL DISTRIBUTION

47. 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.
48. D. Galas, Director, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, Washington, DC 20585.
49. 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.
50. W. Snyder, Department of Radiology, Bowman Gray School of Medicine, 300 South Hawthorne Drive, Winston Salem, NC 27103.
51. G. Featheringham, Cray Research, Inc., Suite 1331N, 1331 Pennsylvania Avenue NW, Washington, DC 20004.
52. M. D. Zorn, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720.
53. J. Fickett, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545.
54. A. Loh, Scientific Applications Marketing Group, Digital Equipment Corporation, Marlboro, MA 01752-9102.
55. C. Watanabe, Genetech, Inc., 1861 Drake Drive, Oakland, CA 94611.
56. C. A. Fields, Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001.

57. E. Branscomb, Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550.
58. A. Lapedes, Center for Human Genome Studies, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545.
59. L. Hood, Division of Biology, California Institute of Technology, Pasadena, CA 91125.
60. T. Hunkapiller, Division of Biology, California Institute of Technology, Pasadena, CA 91125.
61. F. Collins, Howard Hughes Medical Institute, University of Michigan Medical Center, Ann Arbor, MI 48109-0618.
62. E. Lander, Whitehead Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

63-72. Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831.