



3 4456 0366513 9

# ornl

ORNL/TM-12174

**OAK RIDGE  
NATIONAL  
LABORATORY**

**MARTIN MARIETTA**

## **Computer-Based Construction of Gene Models Using the GRAIL Gene Assembly Program**

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

OAK RIDGE NATIONAL LABORATORY  
CENTRAL RESEARCH LIBRARY  
CIRCULATION SECTION  
4100N ROOM 173  
**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.  
UC-1 1989 1 1 1

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) 578-8401, FTS 526-8401.

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-12174

408  
48

Engineering Physics and Mathematics Division

**Computer-Based Construction of Gene Models  
Using the GRAIL Gene Assembly Program**

J.R. Einstein, R.J. Mural<sup>†</sup>, X.Guan, and E.C. Uberbacher

<sup>†</sup>Biology Division

DATE PUBLISHED — September 1992

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 DE-AC05-84OR21400



3 4456 0366513 9



## CONTENTS

ABSTRACT.....	v
1. INTRODUCTION.....	1
2. FIRST-PASS ASSEMBLY.....	2
3. DETECTION AND TREATMENT OF ERROR CONDITIONS.....	5
4. RESULTS.....	7
5. CONCLUSIONS.....	11
REFERENCES.....	13



# 1. INTRODUCTION

A significant long-term challenge facing the Human Genome Project is to develop computer-based technologies to aid in the interpretation of DNA sequence data. A number of general approaches to this problem, including sequence comparison and homology-based methods (1), and pattern recognition using statistical methods (2) and neural networks (3,4,5), have proved valuable in the process of interpreting genomic sequence regions and cDNAs. Recent advances in the characterization of genes by computer-based sequence analysis in such systems as GM (6), Geneid (7), and GRAIL (8,9) have provided real hope that the construction of accurate gene models for most genes found in genomic sequence data will be feasible within the time-frame of the development of high-speed DNA sequencing methods.

GRAIL is a modular expert system being developed to characterize genomic sequences in terms of genes and various control elements. A number of program modules for GRAIL have been constructed and extensively tested. Existing components include the Coding Recognition Module (CRM), which is the heart of the current GRAIL e-mail server system, the Exon Interpretation Rule Base (EIRB), which extracts information such as preferred translation frame and strand for coding regions, the splice-junction Acceptor Recognition Module (ARM) and splice-junction Donor Recognition Module (DRM), the Translation Initiation Module (TIM), and the Gene Assembly Program (GAP). These modules and others are being designed to work within an integrated expert system framework which makes use of a blackboard control architecture.

The research described in this report outlines the methods and performance of the prototype GAP module, which links and scores a variety of features provided by other modules and constructs potential gene models. The purposes of this prototype are to test the effectiveness of the present feature-recognition tools for gene model construction, to further develop principles for automated gene assembly, and to explore error sources and correction strategies. For testing purposes, in the absence of the complete system framework, the GAP prototype has been linked to other modules in a relatively simple manner without making use of the blackboard control architecture.

## 2. FIRST-PASS ASSEMBLY

The GAP program module identifies and assembles gene models on the basis of the following input:

- (a) a genomic sequence known or assumed to contain one gene;
- (b) a coding probability (CP) function over the sequence;
- (c) a probable-translation-frame (PTF) function over the sequence;
- (d) a statistically determined DNA coding strand for the gene;
- (e) the estimated edges of exons;
- (f) the positions and scores of possible splice junctions;
- (f) the positions of possible translation-initiation sites.

The coding probability (CP) function (input b) comes from the Coding Recognition Module (CRM) (8). The version of the CRM used in this study incorporates seven "sensor" algorithms, each designed to provide an indication of the coding potential of a given window (normally 100 bases wide) of sequence. The outputs of the sensor algorithms are input to a neural network having one output node which provides the coding probability score. Through training, the neural network adjusts its internal weights to recognize input sensor values and correlations between sensor values which correspond to coding regions. The Donor and Acceptor Recognition Modules (input (f)) are based on similar sensor-driven neural network systems. Inputs (c) and (d) are determined by a rule-based module called the Exon Interpretation Rule Base or EIRB. This module uses statistical information derived from particular sensors in the CRM to pick the DNA coding strand for the gene, and determine the probable translation frame for each potential exon. At this stage the probable translation frame of each exon is determined independently of other exons to which it may be linked. The EIRB also contains rules to estimate the edges of exons (input (e)) from the size and shape of the coding probability function calculated by the CRM. These edge estimation rules have been determined empirically by statistically examining coding prediction peaks and their relationship to actual exon boundaries. Input (g) for possible translation initiation sites is made statistically using an algorithm similar to that described by Kozak (10).

In the following discussion which describes how GAP assembles putative gene models, the term "coding region" denotes a portion of the sequence identified as a coding exon, the edges of which are splice junction (or translation start or stop codon) positions, intended to be equal or approximately equal to the real edges of

the exon. In short, a “coding region” (CR) is an approximation to an exon being considered and tested in a given gene model.

The procedure that GAP uses to assemble possible genes comprises the following steps:

- (1) An attempt is made to reduce the set of possible translation frames (defined relative to the unspliced sequence) for each CR, on the basis of the PTF function, and the presence or absence of stop codons within the “heart” of the CR (region between the ranges of possible acceptor and donor splice junctions) for each possible translation frame. In other words, certain frames for a given CR are not possible because of stop codons (in those frames) between all possible splice junctions for that CR.
- (2) Sets of possible donor and acceptor splice junctions, taken from input (f), are assigned to each CR, based on their positions relative to initial EIRB-supplied estimated edges of the CR.
- (3) All possible gene models consistent with the input information are assembled and scored. Each gene model consists of a set of CR’s based on the CP function, spliced by a combination of the possible donors and acceptors, subject to the following rules.
  - (a) no stop codons exist (in-frame) within the assembled gene model;
  - (b) exons have a certain minimum length (a program parameter);
  - (c) introns have a certain minimum length (a program parameter).

Furthermore, for a gene model to be accepted, it must have a translation-initiation site, taken from input (g), the position of which satisfies two conditions: being within an acceptable distance from the edge of the leftmost (5’) CR, and in agreement with the translation frame assigned to that CR from the probable-translation-frame function and the splicing procedure.

The assigned score of each completely assembled gene model is the geometric mean of the scores of all utilized splice junctions, times a “translation-frame-discrepancy” score, which is related to the degree of agreement (or disagreement) of the gene model with the PTF function across all CR’s of the model. For purposes of scoring splice junctions, each initial score provided by the ARM or DRM is multiplied by a factor which decreases with distance of the putative junction from its EIRB-predicted coding-region edge. That is, a splice junction with a high initial score because of close coincidence with the consensus sequence receives a lower score if it is far outside the initial edges of the CR or deeply within them.

The end of each gene model is the first stop codon encountered beyond the leftmost (5') edge of the rightmost (3') CR, consistent with the translation frame assigned to this CR by the splicing procedure, provided the minimum exon size criterion is met.

A general problem in assembling genes in the manner outlined above is the large number of splice junction combinations to be considered. For example, in the sequence of HUMPAIA (GenBank: Human Plasminogen Activator Inhibitor-1), containing eight exons, there are potentially about  $3 \times 10^8$  combinations of possible splice junctions. GAP greatly reduces this number by (1) splicing in one direction at a time from the best-determined CR, for one of its possible translation frames; (2) discontinuing splicing beyond any junction which implies a violation of one of the previously described rules; (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. The following examples illustrate the methods by which GAP prunes the search tree. When splicing from left to right (from  $CR_i$  to  $CR_j$ ) the donor positioned to the right of  $CR_i$  is not accepted if there is a stop codon (in the translation frame already assumed for  $CR_i$ ) between the left edge of  $CR_i$  and the donor, or if the implied exon is too short. An acceptor for  $CR_j$  is not accepted if, given the already assumed donor, an "impossible" translation frame is implied for  $CR_j$ , or if the implied intron between  $CR_i$  and  $CR_j$ , is too short. Non-acceptance terminates the search along the current branch of the tree.

By these methods, and because the number of possible translation frames had been reduced to one for several of the CR's, the number of individual splice-junction tests actually done for HUMPAIA was 88300, requiring less than 15 minutes on an IBM PC/AT.

### 3. DETECTION AND TREATMENT OF ERROR CONDITIONS

After first-pass assembly for most sequences in the test set, none of the highest-scoring gene models represents the correct description of the gene, although in many cases the agreement is very close. Possible reasons for the lack of complete agreement obviously include imperfections in the GRAIL modules which provide input to GAP, and limitations in the methods used by GAP. Although efforts to improve all methodologies will continue, it is likely that they will never be perfect, and methods for the detection and correction of errors in initial splicing and assembly will be necessary.

In the current version, GAP detects errors in first-pass assembly in the following three ways:

- (1) no acceptable translation start can be found for the leftmost (5') CR;
- (2) no acceptable splice can be made between a pair of adjacent CR's suitably consistent with the probable-translation-frame function for the two CR's;
- (3) the stop codon for the rightmost CR is unacceptably distant from that CR's right (3') edge, and the calculated PTF function over the region between that CR and the stop codon disagrees to an unacceptable degree with the translation frame assigned to the CR.

At present, error conditions are corrected solely by the insertion or deletion of one or more CR's. In other words, it is currently assumed for each of the three above conditions that the problem is due to missed or false CR's. For example, in case (2) above (no acceptable splice between  $CR_i$  and  $CR_j$ ) there are three possible actions: the insertion of a new CR between  $CR_i$  and  $CR_j$ , the deletion of  $CR_i$ , and the deletion of  $CR_j$ . For case (3), possible actions are the deletion of the rightmost CR, and the insertion of a region beyond it. There are presently five possible actions for dealing with case (1), including deletion of the leftmost CR and insertion of a CR to its left.

The permissibility of an insertion or deletion is governed by several program parameters: the maximum size of an inserted CR (i.e., the assumed maximum size of CR's which could be missing in the CP function), the maximum size of a CR which can be deleted (i.e., the assumed maximum length of a false CR in the CP function), the minimum size of an exon, and the minimum size of an intron.

The insertion of a CR between  $CR_i$  and  $CR_j$  [case (2) above] is accomplished by searching the input splice-junction lists for acceptor-donor pairs which satisfy the programs conditions. More sophisticated methods for evaluating potentially missed exons are in progress.

Each modification [insertion(s), deletion(s), or a combination] is handled independently of others - i.e., the insertions and/or deletions are made in the original set of CR's, and modifications of modifications are not considered. Attempts at splicing take place, and the resulting gene models of sufficiently high score are saved. Insertions are currently considered without reference to the CP function in the relevant region; however, new sensors for the CRM which are in development should make it possible to zero in on missing regions more reliably in the near future.

Without insertion and deletion of CR's (first-pass assembly), GAP was unable to obtain complete gene models for a number of sequences in the test set, because of its inability to find acceptable translation-frame consistent splices between a pair of adjacent CR's, or to locate an acceptable translation-initiation site. In addition to allowing the completion of gene models for many sequences, the gene assembly process including error detection and correction resulted in a net improvement in gene models compared to the results of the GRAIL CRM alone - see "Results" section.

## 4. RESULTS

The results of gene assembly using GAP and the other GRAIL modules are summarized in Table 1a,b. The table presents results for the portions of genes recognized using the CRM alone (equivalent to the e-mail GRAIL server system) in comparison to the assembled gene models constructed with GAP. Only the highest-scoring gene model for each sequence was used for the statistics. Percentages for both the "found" and "false" categories are referenced to the true (experimentally determined) numbers of exons and bases.

For both exons (Table 1a) and bases (Table 1b), the assembly process provides significant benefit in terms of the accuracy of the gene models compared to the initial predictions of the CRM alone. Insertion/deletion of CR's is responsible for three of the perfect gene models, which could not be obtained in the first assembly pass. Insertion/deletion is also responsible for finding a number of CR's missed by the CRM, and for deleting a portion of the false positive CR's found by the CRM. Of 132 real exons in the 26 sequences, the CRM found 106, missed 26, and also gave 18 false positives (the statistics for large exons ( $> 100$  bases) alone are considerably better - see Table 1a). GAP inserts 5 of the 25 missed CR's, and deletes 4 of the 18 false positive CR's. Especially significant is the improvement for the shortest exons ( $< 50$  bases), where 60% of these, originally missed by the CRM, are recovered through insertion during assembly! Also, a large percentage of false positive coding regions, which usually show up in the CP function as very short CR's are eliminated because they could not be spliced, etc. For example, of the original eight false positive very short exons ( $< 25$  bases), GAP is able to remove seven, while inserting three real exons in the same size category. This result provides evidence that such short exons may be reliably found in genes as the current prototype methods are improved. Because of the complexity of GAP's actions, certain detrimental effects also take place: one large exon is deleted and three false positives created.

The performance in terms of bases shows trends similar to that for exons. A significant percentage of the bases in false positive short CR's is eliminated by error correction/deletion and a significant percentage of bases in real short exons, initially not recognized, is recovered by insertion. On average, the best (highest-scoring) model recovers about 85% of the genetic message, with about 7% false information. Assembly results in a dramatic improvement in recovering the approximate exon edges, compared to the CRM, which generally underestimates exon size. The percentage of correct bases found through assembly is about 20% higher than in the original CRM prediction, mostly due to edge effects in the CRM and the additional exons which were correctly inserted in assembly.

For six of the 26 sequences, a perfect gene model was among the ten highest scoring models. The perfect model placed first in three cases and third, fourth, and eighth in three others. In most cases for which all the exons of a gene have been found and where there are no false positive exons, a small percentage of bases is missing or false. This occurs primarily because a number of incorrect splice junctions are given very high scores by the ARM and DRM. Efforts are underway to improve the performance of these modules.

It should be pointed out that these results are preliminary, using first guess estimates for key parameters. We anticipate that complex methods will be required for dealing accurately with errors in the gene assembly process. An expert system with blackboard control architecture should facilitate the complex reasoning necessary for the adoption of different strategies for different unpredictable circumstances.

**Table 1a**  
**Statistics for Assembly (Single Gene Assumption)**  
**- 26 Genes -**

<b>EXONS</b>									
<b>size</b>	<b>no.</b>	<b>CRM</b>		<b>gap</b>		<b>CRM</b>		<b>gap</b>	
		<b>found</b>	<b>%</b>	<b>found</b>	<b>%</b>	<b>false</b>	<b>%</b>	<b>false</b>	<b>%</b>
1- 24	3	0	0	3	100	8	267	1	33
25- 49	7	1	14	3	43	5	71	5	71
50- 74	11	6	55	7	64	4	36	4	36
75- 99	16	11	69	11	69	1	6	3	19
100-124	21	18	86	18	86	0	0	0	0
125-149	21	20	95	19	90	0	0	1	5
150-174	16	16	100	16	100	0	0	0	0
175-199	12	11	92	11	92	0	0	0	0
>199	25	23	92	23	92	0	0	0	0
small	37	18	49	24	65	18	49	13	35
big	95	88	93	87	92	0	0	1	1
all	132	106	80	111	84	18	14	14	11

**Table 1b**  
**Statistics for Assembly (Single Gene Assumption)**  
**– 26 Genes –**

<b>BASES</b>									
		<b>CRM</b>		<b>gap</b>		<b>CRM</b>		<b>gap</b>	
<b>size</b>	<b>no.</b>	<b>found</b>	<b>%</b>	<b>found</b>	<b>%</b>	<b>false</b>	<b>%</b>	<b>false</b>	<b>%</b>
1– 24	47	0	0	47	100	139	296	15	32
25– 49	226	22	10	108	48	200	88	159	70
50– 74	695	121	17	410	59	335	48	502	72
75– 99	1408	459	33	964	68	140	10	298	21
100–124	2357	1231	52	1904	81	57	2	53	2
125–149	2871	1290	45	2262	79	21	1	195	7
150–174	2623	1729	66	2466	94	68	3	50	2
175–199	2239	1296	58	1746	78	52	2	37	2
>199	10169	8089	80	9285	91	137	1	193	2
small	2376	602	25	1529	64	814	34	974	41
big	20259	13635	67	17663	87	335	2	528	3
all	22635	14237	63	19192	85	1149	5	1502	7

## 5. CONCLUSIONS

The results of the prototype GAP module of GRAIL for 26 sequences demonstrate that most of the coding regions of a DNA sequence may be found using an automated procedure based on computer analysis of the sequence itself. There is reason to expect that further development of the system will provide progressively more accurate results.

Further developments in progress or planned include the following:

- (1) Improvement in the basic recognition technologies employed in the CRM, ARM, DRM, etc., which provide putative feature input to the assembly process. In particular, improvements in the Coding Recognition Module by inclusion of other sensors, including high-order Markov analysis, show promise for more correctly identifying short exons which are otherwise missed.
- (2) Improvement of the present GRAIL system methodologies, in particular, through the incorporation of the modules into a blackboard expert system framework.
- (3) Methods to better identify the limits of individual genes, so that regions with multiple genes can be dissected and single genes assembled separately.
- (4) Implementation of GAP on a parallel computer in order to facilitate the analysis of genes having large numbers of exons.



## REFERENCES

1. M.D. Adams, M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C.Fields and J.C. Vantor, *Nature* 355, 632 (1991).
2. R. Staden, *Nucl. Acids Res.* 12, 505-519 (1984).
3. A. Lapedes, C. Barnes, C. Burks, R. Farber and K. Sirotkin, in "Computers and DNA", *SFI Studies in the Science of Complexity*, G. Bell and T. Marr, eds., Vol. VII, 157-182, Addison-Wesley (1989).
4. S. Brunak, J. Engelgrecht and S. Knudsen, *Nucl. Acids. Res.* 18, 4797-4801 (1990).
5. S. Brunak, J. Engelbrecht and S. Knudsen, *J. Mol. Biol.* 220, 49- 65 (1991).
6. C. Fields and C.A. Soderlund, *CABIOS* 6, 263-270 (1990).
7. R. Guigo, S. Knudsen, N. Drake and T. Smith. *J. Mol. Biol.* 226, 141-157 (1992).
8. E. Uberbacher and R.J. Mural, *Proc. Natl. Acad. Sci. USA* 88, 11261-11265 (1991).
9. X. Guan, R.J. Mural, J.R. Einstein, R.C. Mann, E.C. Uberbacher, *Proceedings of the 8th IEEE Conference of Artificial Intelligence Applications*, Monterey, CA, March 2-6 (1992).
10. M. Kozak, *J. Cell Biol.* 108, 229-241 (1989).



## INTERNAL DISTRIBUTION

- |                         |                                 |
|-------------------------|---------------------------------|
| 1. B. R. Appleton       | 36. R. W. Brockett (consultant) |
| 2-6. J. R. Einstein     | 37. J. J. Dorning (consultant)  |
| 7. N. Wright Grady      | 38. J. E. Leiss (consultant)    |
| 8-12. X. Guan           | 39. N. Moray (consultant)       |
| 13. R. E. Hand, Jr.     | 40. M. F. Wheeler (consultant)  |
| 14. F. C. Hartman       | 41. EP&MD Reports Office        |
| 15. J. P. Jones         | 42-43. Laboratory Records       |
| 16-20. R. C. Mann       | Department                      |
| 21. T. J. Mitchell      | 44. Laboratory Records,         |
| 22-26. R. J. Mural      | ORNL-RC                         |
| 27. C. E. Oliver        | 45. Document Reference          |
| 28. D. E. Reichle       | Section                         |
| 29. C. R. Richmond      | 46. Central Research Library    |
| 30-34. E. C. Uberbacher | 47. ORNL Patent Section         |
| 35. R. C. Ward          |                                 |

## EXTERNAL DISTRIBUTION

48. 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
49. D. Galas, Director, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, Washington, DC 20585
50. B. J. Barnhart, Program Manager, Human Genome Project, Office of Health and Environmental Research, Office of Energy Research, U.S. Department of Energy, Washington, DC 20585
51. W. Snyder, Department of Radiology, Bowman Gray School of Medicine, 300 South Hawthorne Drive, Winston Salem, NC 27103
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. C. Watanabe, Genetech, Inc., 1861 Drake Drive, Oakland, CA 94611
55. C. A. Fields, Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001
56. E. Branscomb, Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
57. A. Lapedes, Center for Human Genome Studies, Los Alamos National Laboratory, P.O. Box 1663, Las Alamos, NM 87545
58. T. Hunkapiller, Division of Biology, California Institute of Technology, Pasadena, CA 91125
59. F. Collins, Howard Hughes Medical Institute, University of Michigan Medical Center, Ann Arbor, MI 48109-0618
60. E. Lander, Whitehead Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139
- 61-70. Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831

71. Yutaka Akiyama, Ph.D., Associate Professor, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, JAPAN
72. Dr. John C. Wooley, Deputy Associate Director/OHER, ER71, Department of Energy, Washington, DC 20585