MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0422643 3

ORNL/TM-12976

OAK RIDGE NATIONAL LABORATORY

MARTIN MARIETTA

Alignment of DNA and Protein Sequences Containing Frameshift Errors

X. Guan
E. C. Uberbacher

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.
MARTIN MARIETTA

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-12976

Computer Science and Mathematics Division

Alignment of DNA and Protein Sequences Containing Frameshift Errors

X. Guan and E. C. Uberbacher

DATE PUBLISHED - April 1995

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 0422643 3

CONTENTS

	Page No.
ABSTRACT	v
1. Introduction	1
2. Methods	2
3. Results and Discussions	6
3.1. Binary comparison and alignment	9
3.2. Comparison to protein family and the database	9
4. Conclusion	13
5. Acknowledgements	16
References	17

ABSTRACT

Molecular sequences, like all experimental data, are subject to error. Many current DNA sequencing protocols have very significant error rates and often generate artifactual insertions and deletions of bases (indels) which corrupt the translation of sequences and compromise the detection of protein homologies. The impact of these errors on the utility of molecular sequence data is dependent on the analytic technique used to interpret the data. In the presence of frameshift errors, standard algorithms using six frame translation can miss important homologies because only sub-fragments of the correct translation are available in any given frame. We present a new algorithm which can detect and correct frameshift errors in DNA sequences during comparison of translated sequences with protein sequences in the databases. This algorithm can recognize homologous proteins sharing 30% identity even in the presence of a 7% frameshift error rate. Our algorithm uses dynamic programming, producing a guaranteed optimal alignment in the presence of frameshifts, and has a sensitivity equivalent to Smith-Waterman. The computational efficiency of the algorithm is $O(nm)$ where n and m are the sizes of two sequences being compared. The algorithm does not rely on prior knowledge or heuristic rules and performs significantly better than any previously reported method.

1. Introduction

Knowledge of protein and nucleic acid sequences is central to many aspects of modern biology. Experimentally determined DNA sequences are generally used as the basis for determining the sequence of amino acids in encoded proteins using multiple-frame translation or pattern recognition methods. At present, comparison among the sequences is most commonly done using protein sequence (i.e., deduced amino acid sequence) to understand functional relationships among proteins in the same and in different species.

A number of sequence comparison algorithms have been introduced. For rigorous analysis, the dynamic programming algorithm proposed by Smith and Waterman [8], and modified by Gotoh [3], can be used to find the optimal alignment of two sequences. For fast identification of homologous sequences, heuristic methods such as BLAST [1], based on a probabilistic model, and FastA [4], based on initial k-tuple matches, can be used. These heuristic methods are not as sensitive or accurate as the full dynamic programming method.

As experimental data, sequences are subject to error. Error can arise from a variety of sources such as migration artifacts in sequencing gels, improper gel reading by computer software, methods for removing inconsistencies during sequence assembly, and in data entry. Although some level of error in sequences seems inevitable, different sequencing strategies have different intrinsic error rates and types. Some approaches such as single pass cDNA sequencing or cosmid skimming have virtually no redundancy checks on the sequence, and can have error rates in excess of 10%, while high-redundancy (shotgun) genomic sequencing can be about 1% or less.

In considering the impact of errors in nucleotide sequences, it is useful to consider not only the frequency of errors but also their type (substitution, deletion, or insertion of one or more bases). In particular, derived amino acid sequences are very differently affected by errors of different types. The most common type of error involves an incorrect read for a base which results in a substitution of one base for another. This weakens the recognition of homologies but in a linear and fairly gentle way. The requirement of a consistent reading frame for translation, free of insertions or deletions (i.e., without "frameshift" errors), is a much more stringent constraint on the quality of sequence data which will be used to recognize homologies with standard sequence comparison algorithms. In the presence of insertion-deletion errors, resulting translations for the coding regions may have only certain parts correct in any one frame. Therefore, when comparing the protein translation to sequences in a protein database, useful homologies can be easily lost.

The trend toward efficient sequencing methods which inherently have less redundancy creates a need for systems which can detect coding regions and detect sequence homologies in the presence of significant error rates. Several methods have been developed to deal with sequence errors when comparing sequences. Posfai and Roberts [5] used an approach based on FastA to compare the 6-frame translations of a DNA sequence to sequences in a protein database. In this method, if two

strongly matched segments are adjacent and are in different reading frames, a possible frameshift site has been found. States and Botstein [7] proposed an algorithm based on a probabilistic model incorporating codon usage information and a non-uniform distribution of error probabilities. In their algorithm, a DNA sequence which may have errors is compared with a protein sequence, and uncertainties are assigned for each base in the DNA sequence as well as for an insertion or a deletion at that base using the prior information. In this method, and for a specific organism, proteins with ≥ 33% sequence identity can be recognized in the presence of 1% frameshifting errors (deletion or insertion) and 5% base substitution. Both existing methods either rely on a probability model or on heuristic rules and involve methods with reduced sensitivity and accuracy compared to more robust dynamic programming methods.

While these algorithms have some specialized utility, we describe here a more general and more sensitive approach to dealing with errors in the course of sequence comparison. We have previously described a coding recognition system which functions in the presence of indels [9], and in this paper address methods for recognition of protein sequence homologies when frameshifts are present. The algorithm can recognize homology between DNA sequence translations and related protein sequences in the presence of significant (several percent) indel rates in DNA sequences. Furthermore, in this method and for each identified homology, the frameshift sites in the query sequence are located as well. The method is very general because it does not use heuristic rules, specific codon usage information, or any knowledge other than DNA sequences themselves. Our algorithm uses a rigorous dynamic programming approach similar to the Smith-Waterman algorithm and is therefore more sensitive and accurate than other current methods.

2. Methods

The Smith-Waterman algorithm for local alignment modified by Gotoh [3] is as follows. Given two sequences $A = a_1a_2..a_m$ and $B = b_1b_2..b_n$, a cost function, w, for k insertions or deletions is defined as $w(k) = -(u*k+v)$, where $u \geq 0$ and $v \geq 0$. The alignment of A and B is performed by creating a socre matrix D whose values are computed systematically from the upper left corner of the matrix to the lower right. Two other matrices P and Q are used in calculating matrix D.

$$D(i,j)=\max\{D(i-1,j-1)+c(b_i,a_j),\ P(i,J),\ Q(i,j),\ 0\}$$
$$0<i\leq m,\ 0<j\leq n$$

where

$$P(i,j)=\max\ \{D(i-1,j)+w(1),\ P(i-1,j)+u\}$$

and

$$Q(i,j)=\max\ \{D(i,j-1+w(1),\ Q(ij-1)+u\}$$

$$P(0,k)=Q(k,0)=D(k,0)=D(0,k)=0,\ \forall k>0$$

$(b_i,\ a_j)$ represents the similarity between a_i and b_j where

$$c(b_i,a_j)=\begin{cases} \leq 0 & \text{if } b_i \neq a_j \\ \geq 0 & \text{if } b_i = a_j \end{cases}$$

To find the best local alignment, the maximum score in the score matrix D is found first, which corresponds to the right end of an optimal local alignment, then a traceback procedure is performed to find the left end of the optimal local alignment. An example is given in Figure 1.

When a DNA sequence with frameshift errors is translated into protein sequence and then aligned with sequences in a protein database, correctly translated segments will fall into different frames. In the comparison for any given frame, segments will be incorrect, resulting in significant mismatch, even if the two sequences are truly homologous. An example is shown in ***Figure 2***. The rat neutrophil elastase mRNA and human cercarial elastase share 26% sequence identity. The translation of rat mRNA is first aligned with human neutrophil elastase using the standard Smith-Waterman algorithm to show the homology. When three deletion errors are introduced in the rat mRNA sequence, and a three-frame comparison is made using Smith-Waterman, the best local homology found in any frame is very poor (***Figure 2b***). In a comprehensive database search, the original alignment was at position 219 in the result list, but with the indels the best local alignment moves down to position 1410, effectively becoming lost in the noise.

As in the above example, consider what happens when the translation in all three frames is aligned independently. Suppose the coding region starts in frame 1, and a deletion occurs in the sequence. The frame 1 translation is aligned with the protein sequence until the deletion site, then the frame 2 translation is aligned with the protein sequence. If a second deletion is encountered, the frame 3 translation picks up the alignment at this site. If instead an insertion were to occur after the first deletion, the alignment would return to the frame 1 translation. If we align the three frame translations independently, segments of alignments will be scattered throughout the three score matrices.

a. The score matrix D.

		M	C	T	G	S	L	V	S	T	R	A	V	L	T	A	G	H
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	7	6	0	0	0	0	0	0	0	0	0	0	0	6	0
G	0	0	0	0	13	6	1	0	0	0	0	0	0	0	0	0	6	4
T	0	0	0	5	2	14	5	2	1	5	0	0	0	0	5	0	0	4
S	0	0	0	1	5	6	12	3	6	2	4	1	0	0	1	6	0	0
L	0	2	0	0	0	3	10	13	2	5	0	3	2	4	0	0	2	0
I	0	1	1	0	0	1	5	13	11	1	2	0	6	4	3	0	0	0
N	0	0	0	1	0	1	0	2	14	11	2	1	0	3	4	1	0	1
D	0	0	0	0	0	0	0	1	3	13	9	1	0	0	2	2	0	0
Q	0	0	0	0	0	0	0	0	2	2	14	8	2	1	0	1	0	0
W	0	0	0	0	0	0	0	0	1	1	3	11	5	0	0	0	0	0
V	0	1	0	0	0	0	1	4	0	1	2	3	15	6	3	2	1	0
V	0	1	0	0	0	0	1	5	2	0	1	2	7	16	6	4	3	2
C	0	0	10	0	0	0	0	0	4	1	0	1	3	6	15	6	3	2
P	0	0	0	9	0	0	0	0	0	3	0	0	2	4	5	14	4	2
T	0	0	0	5	7	1	0	0	1	5	2	0	1	3	9	5	12	2

b. The best local alignment of the two sequences.

```
A: 3 CGGTSLINDQWVV 15
       : : ::.. . :.
B: 2 CTG-SLVSTRAVL 13
```

Figure 1. Two protein sequences A = HFCGGTSLINDQWVVCPT and b = MCTGSLVSTRAVLTAGH are compared using the Smith-Waterman algorithm. Here w(k) = -(1 * k + 10), and similarities between amino acids are defined by the protein similarity matrix BLOSUM 62 [11]. The maximum score in the score matrix D (see a) is 16 at (15,13) and the corresponding alignment is given in b.

a. The local alignment of the rat mRNA sequence with cercarial elastase.

Identities: 63/234 (27%)

```
Query:   36 PYQVSLNSGYHFCGGSLINDQWVVSAAHCYKSRIQVRLGEHNINVLEGDEQFI----NAA 91
            :.   : .    : :::.. . :..: ::  : . : .    . .  ::.: :    .
Sbjct:   41 PFIAFLTTERTMCTGSLVSTRAVLTAGHCVCSPLPV-IRVSFLTLRNGDQQGIHHQPSGV 99

Query:   92 KIIK--HPNY-------S-SWTLNN-DIMLIKLSSPVKLNARVAPVAL--PSACAPAGTQ 138
            :.     :.          . ::.  :: .. :.  : : . .  ..:  ::   : ::
Sbjct:  100 KVAPGYMPSCMSARQRRPIAQTLSGFDIAIVMLAQMVNLQSGIRVISLPQPSDIPPPGTG 159

Query:  139 CLISGWG---NTLSNGVNNPDLLQCVDAPVLSQADCEAAYPGEITSSMICVGFLEG-GKD 194
              : :.:   :       :  .::.   : ..   .:  :  :    . :::   .  :.
Sbjct:  160 VFIVGYGRDDNDRDPSRKNGGILKKGRATIM---ECRHATNG----NPICVKAGQNFGQL 212

Query:  195 SCQGDSGGPVV--CNGQLQGIVSWGYGCALPDNPG--V-YTKVCNFVGWIQDTI 243
               :::::::..    : . :.::  .:  ::. :   : :  :   . ...  :
Sbjct:  213 PAPGDSGGPLLPSLQGPVLGVVS--HGVTLPNLPDIIVEYASVARMLDFVRSNI 264
```

b. The three frame local alignment of the rat mRNA sequence containing three deletion errors with cercarial elastase. Only small segments from the error-free alignment were picked up in each of the three frame alignments.

frame1:

```
Query:   36 PYQVSLNSGYHFCGGSLINDQWVVSAAH 63
            :.   : .    : :::.. . :..: :
Sbjct:   41 PFIAFLTTERTMCTGSLVSTRAVLTAGH 68
```

frame2:

```
Query:  191 GKDSCQGDSGGP 202
            :.    ::::::
Sbjct:  210 GQLPAPGDSGGP 221
```

frame3:

```
Query:  106 DIMLIKLSSPVKLNARVAPVALP 128
            :: .. :.  : : . .  ..::
Sbjct:  126 DIAIVMLAQMVNLQSGIRVISLP 148
```

Figure 2. The alignments of rat mRNA and cercarial elastase using Smith-Waterman algorithm before and after three deletion errors were introduced.

To find potential frameshift sites, one can examine the segments of alignments in all three frames and look for cases where two segments are adjacent and are in different frames. One might consider heuristic rules to decide which segments to bring into an alignment, especially if the homology is quite strong. However, if the homology is weak, or if there are many such segments, it is difficult to find the overall best alignment because of the complexity and combinatorics involved.

We describe a new algorithm to find the optimal alignment in such a situation and which also locates the indel sites. Since any combination of deletions and insertions places the correct reading frame into one of three frames, no combinatorial calculation is necessary and a computationally efficient dynamic programming approach can be used.

In the standard Smith-Waterman algorithm, a score matrix cell *(i, j)*'s value can depend on three other matrix cells *(i-1, j-1)*, *(i, j-1)* and *(i-1, j)*. In the new alignment algorithm we consider not only the three cells in the same matrix, but also the same three cells in the other two frames' matrices. That is, when computing the alignment for a given frame of translation, we also considered whether there is a better partial alignment in either of the other frames prior to this point that can be continued by shifting the frame to the one under consideration (see ***Figure 3***).

We do this for all three score matrices. When the calculation is completed, segments of matching sequences in the three score matrices are threaded together into sequence alignments. To prevent an alignment from being constructed from random matches from three frames and very frequent frameshifts, a penalty is imposed for shifting the frame. Along with each alignment path we record the frame of the current match. In the traceback procedure, we start from the matrix cell of maximum value and trace back the alignment through the three frames' matrices. If there is true homology between the two sequences aligned, the three score matrices will show basically the same alignment with slight differences at the ends. ***Figure 4*** shows the alignment of the same sequences in ***Figure 2*** using the new algorithm.

3. Results and Discussion

We evaluated the sensitivity and accuracy of our algorithm in two different ways: (1) Using several different examples, a DNA sequence coding region was subject to different error rates and compared with a homologous protein sequence. As a control, the results were compared with that of the randomly permitted sequences of the same base composition. The goal was to evaluate the impact of different error rates on a known sequence alignments, and determine how much error was tolerable before the homology would become insignificant relative to the noise. (2) We searched a DNA sequence with different rates of indel error against a protein database to see whether the sequences from the same protein family were recognized by the algorithm, and how both strong and weak homologies within the result list were affected by different rates of sequence error.

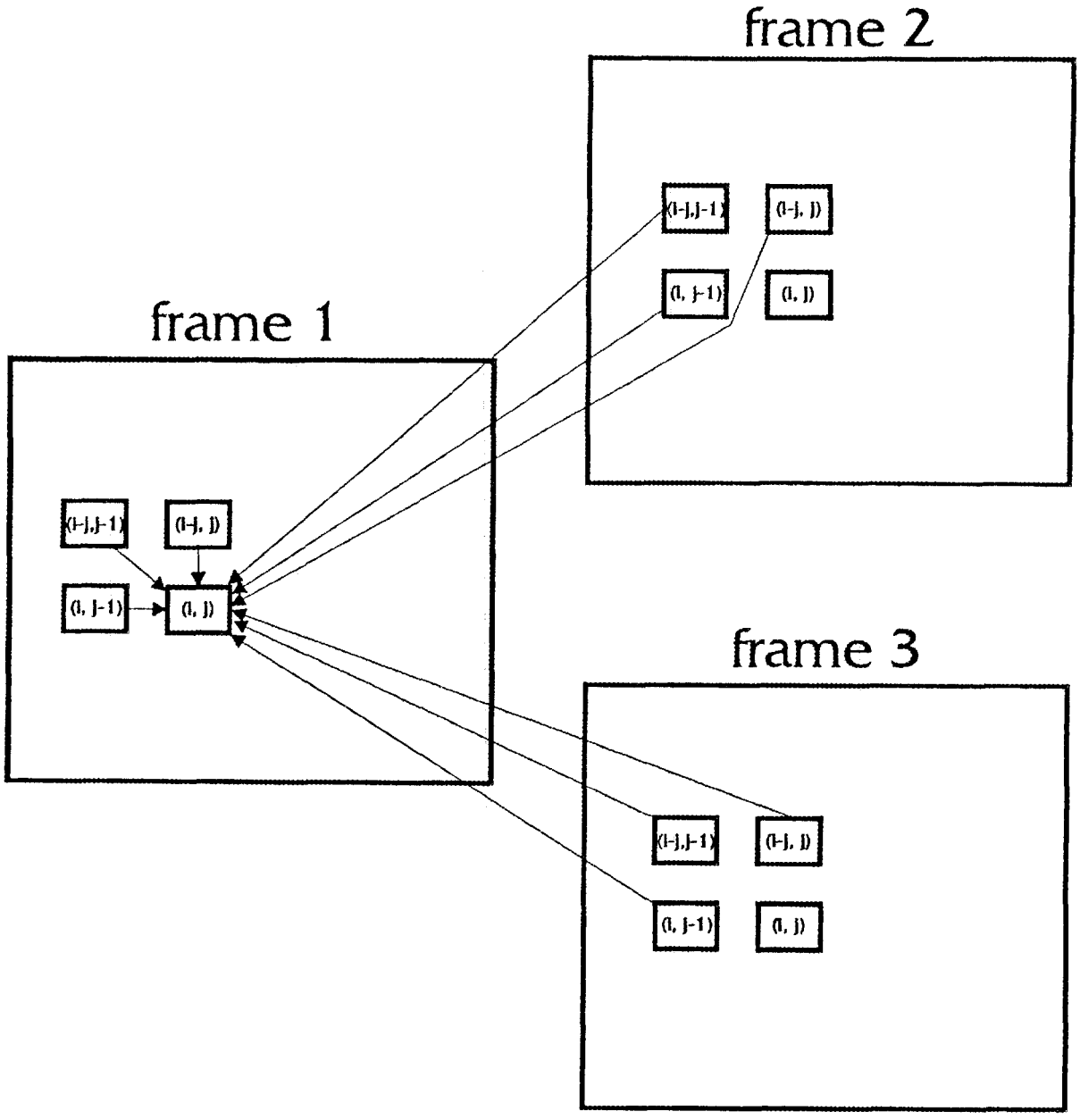


Figure 3. This picture shows the basic idea of our algorithm. When calculating score matrix cell (i, j) for frame 1, we not only consider the three matrix cells (i-1, j-1), (i-1, j), and (i, j-1) in frame 1, but also consider the same three matrix cells in frame 2 and 3, thereby connecting segments of matching sequences in all three frames. This is done for all three score matrices.

```
Identities: 61/234 (26%)

Query:   36 PYQVSLNSGYHFCGGSLINDQWVVSAAHAT-NPIQVRLGEHNINVLEGDEQFI----NAA 90
            :.   : .    : ::::.. . :..: :   .:. : .    . .  ::.: :    .
Sbjct:   41 PFIAFLTTERTMCTGSLVSTRAVLTAGHCVCSPLPV-IRVSFLTLRNGDQQGIHHQPSGV 99

Frame:   41 111111111111111111111111111111133333333333333333333333333333 99

Query:   91 KIIK--HPNY-------S-SWTLNN-DIMLIKLSSPVKLNARVAPVAL--PSACAPAGLS 137
            :.     :.          . ::.  :: .. :.  : : . .  ..:  ::   : :
Sbjct:  100 KVAPGYMPSCMSARQRRPIAQTLSGFDIAIVMLAQMVNLQSGIRVISLPQPSDIPPPGTG 159

Frame:  100 333333333333333333333333333333333333333333333333333333333333 159

Query:  138 ALISGWG---NTLSNGVNNPDLLQCVDAPVLSQADCEAAYPGEITSSMICVGFLEG-GKD 193
              : :::   :        :  .:.   : ..   .:  :  :    . :::    .  :.
Sbjct:  160 VFIVGYGRDDNDRDPSRKNGGILKKGRATIM---ECRHATNG----NPICVKAGQNFGQL 212

Frame:  160 322222222222222222222222222222222222222222222222222222222222 212

Query:  194 SCQGDSGGPVV--CNGQLQGIVSWGYGCALPDNPG--V-YTKVCNFVGWIQDTI 242
               :::::::..    : . :.::  .:  ::. :   : :  :   . ...  :
Sbjct:  213 PAPGDSGGPLLPSLQGPVLGVVS--HGVTLPNLPDIIVEYASVARMLDFVRSNI 264

Frame:  213 222222221111111111111111111111111111111111111111111111 264
```

Figure 4. The alignment of the rat nRNA containing three deletion errors with cercarial elastase using our algorithm.

3.1. Binary comparison and alignment

This evaluation is similar to one performed by States and Botstein [7]. We selected the same two sequences, rat trypsin and human neutrophil, which share 33% sequence identity. We first tested the impact of substitution errors on the sequence alignment. For each error rate, 100 runs of the alignments were performed, each with a new set of errors randomly introduced in the coding region of the sequence. The distribution of the scores is plotted in ***Figure 5***. As a control for the comparison, the DNA sequence was randomly permutted, translated, and compared with the target protein sequence. The score distribution is also plotted on the same graph in ***Figure 5***. The purpose was to examine the discrimination between the true alignment and the resulting random alignments.

As can be seen in ***Figure 5***, as more errors were introduced, the alignment and discrimination becomes weaker (as expected). However the discrimination remains significant to very high substitution rates. The performance is basically the same as standard Smith-Waterman in this case (base substitution but no indels). With 33% sequence identity between the two sequences, the algorithm can recognize the homology in the presence of up to 15% substitution errors, as opposed to about 5% for the States and Botstein algorithm.

A similar evaluation was done to measure the effects of indel errors. Indel errors cause more damage to sequence alignment than substitution errors because indels produce frameshifts while substitutions only reduce sequence identity or similarity. For the same test sequences, the algorithm adequately separated the actual homologies from noise in the presence of a 7% frameshift error rate (see ***Figure 6***), compared to about 1% for the States and Botstein algorithm. As we calculate the three score matrices, the frameshift information is recorded and can be used during traceback to get the positions of the frameshifts which are included in the final alignment (see ***Figure 4***).

3.2. Comparison to protein family and the database

One way of evaluating a protein sequence comparison algorithm is to examine how well a member of a protein family hits the other members of the family, including those distantly related. As an example of this we use the Desulfovibrio vulgaris (Hildenborough) gene for cytochrome c3 as the query sequence. We first used the standard Smith-Waterman algorithm to find the homologous sequences in Swissprot protein database ((version 30) Bairoch and Boeckmann [10]). These matching sequences share sequence identity ranging from 88% to 30% with the query. The result is shown in ***Figure 7a***. The new algorithm was tested using several rates of indel in the query sequence. As is shown in ***Figure 7***, the algorithm functions well with a 5% frameshift rate, still recognized the homologous sequences in the top 10 of the returned sequence list.

An issue that often accompanies methods based on Smith-Waterman algorithm is their efficiency. The basic form of our algorithm is a factor of about seven times

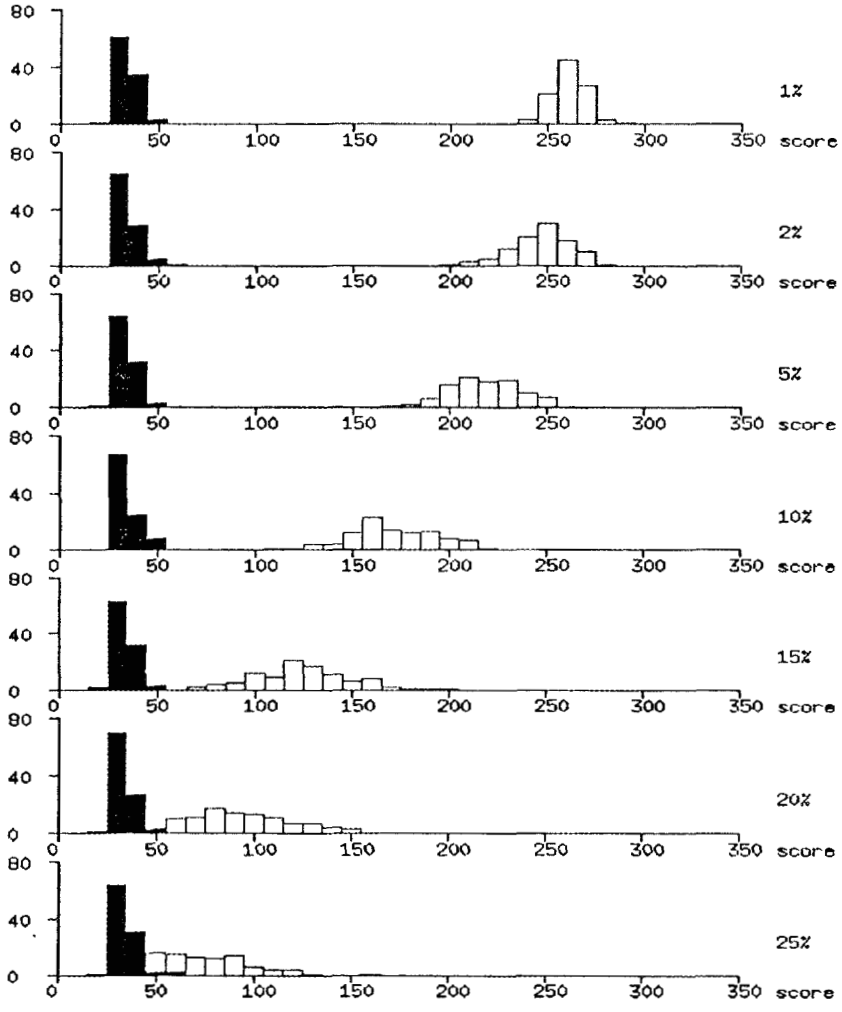


Figure 5. Two sequences, rat trypsin and human neutrophil which share 33% sequence identity, were compared with a range of substitution error rates (1% - 25%). For each error rate, 100 runs of the alignments, each with a new set of errors introduced in the sequences, were performed and the distribution of the alignment scores were plotted as the open bars. The solid bars represented the distribution of the alignments scores of the same DNA sequences, randomly permuted, translated, and compared with the target protein sequence.

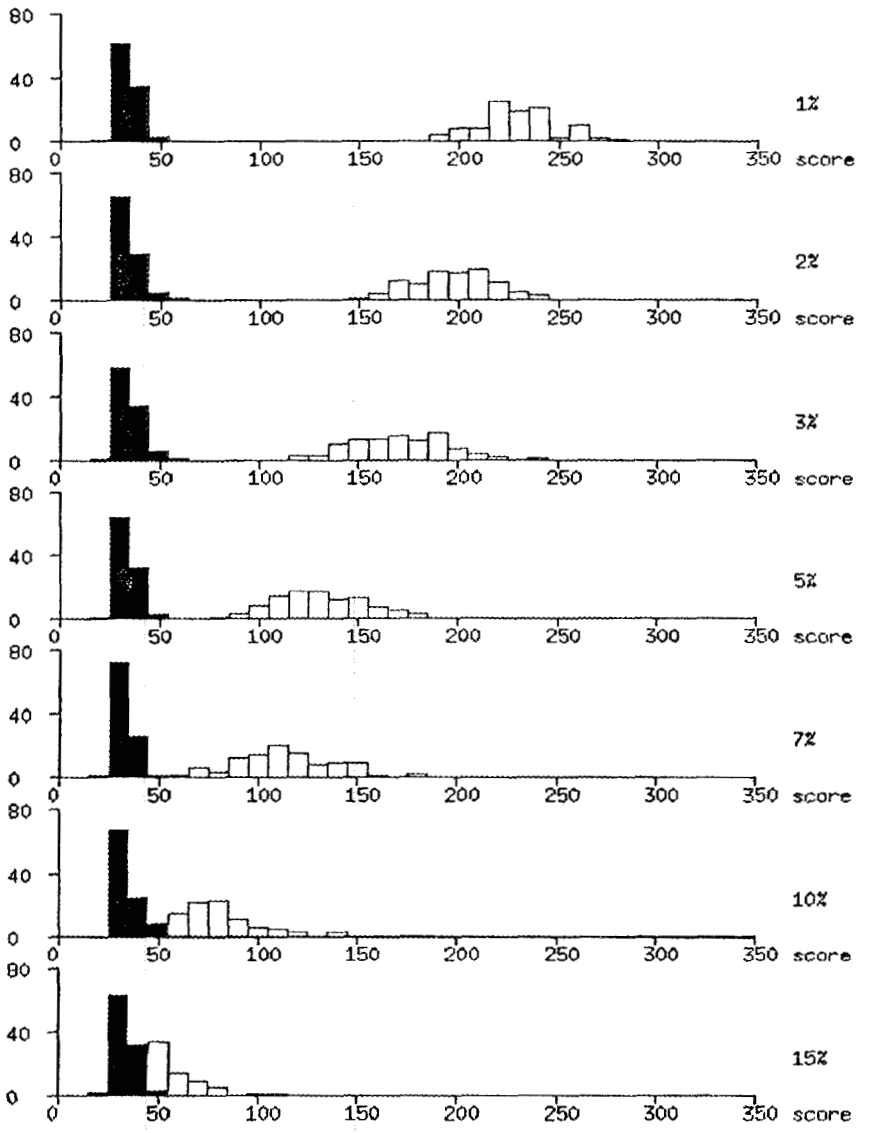


Figure 6. Discrimination for sequences with indel errors. Indel error rates (1% - 15%) were introduced and the distributions of alignments scores were plotted as in Figure 5.

a. Top 10 sequences returned by the standard Smith-Waterman algorithm. These sequences share sequence identity with cyc3de ranging from 88% to 31%.

1	712	P00131	CYC3_DESVH	DESULFOVIBRIO VULGARI
2	543	P00132	CYC3_DESVM	DESULFOVIBRIO VULGARI
3	266	P00133	CYC3_DESGI	DESULFOVIBRIO GIGAS.
4	160	P00135	CYC3_DESSA	DESULFOVIBRIO SALEXIG
5	151	P00134	CYC3_DESDE	DESULFOVIBRIO DESULFU
6	140	P38554	CY32_DESDN	DESULFOVIBRIO DESULFU
7	123	P00136	CY31_DESDN	DESULFOVIBRIO DESULFU
8	75	P24092	HMWC_DESVH	DESULFOVIBRIO VULGARI
9	59	P32707	NRFB_ECOLI	ESCHERICHIA COLI.
10	58	P24735	AMPC_PSEAE	PSEUDOMONAS AERUGINOS

b. Top 10 sequences returned by our algorithm when 1% indels were added.

1	659	P00131	CYC3_DESVH	DESULFOVIBRIO VULGARI
2	523	P00132	CYC3_DESVM	DESULFOVIBRIO VULGARI
3	251	P00133	CYC3_DESGI	DESULFOVIBRIO GIGAS.<
4	153	P00134	CYC3_DESDE	DESULFOVIBRIO DESULFU
5	152	P00135	CYC3_DESSA	DESULFOVIBRIO SALEXIG
6	143	P38554	CY32_DESDN	DESULFOVIBRIO DESULFU
7	128	P00136	CY31_DESDN	DESULFOVIBRIO DESULFU
8	97	Q02817	MUC2_HUMAN	HOMO SAPIENS (HUMAN).
9	89	P13728	SGS3_DROYA	DROSOPHILA YAKUBA (FR
10	87	Q06885	GP10_DICDI	DICTYOSTELIUM DISCOID

c. Top 10 sequences returned by our algorithm when 5% indels were added.

1	414	P00131	CYC3_DESVH	DESULFOVIBRIO VULGARI
2	332	P00132	CYC3_DESVM	DESULFOVIBRIO VULGARI
3	154	P00133	CYC3_DESGI	DESULFOVIBRIO GIGAS.<
4	99	P00134	CYC3_DESDE	DESULFOVIBRIO DESULFU
5	96	P38554	CY32_DESDN	DESULFOVIBRIO DESULFU
6	90	P00135	CYC3_DESSA	DESULFOVIBRIO SALEXIG
7	75	P00136	CY31_DESDN	DESULFOVIBRIO DESULFU
8	74	Q03391	NME4_MOUSE	MUS MUSCULUS (MOUSE).
9	69	P13730	SGS3_DROER	DROSOPHILA ERECTA (FR
10	68	P13728	SGS3_DROYA	DROSOPHILA YAKUBA (FR

Figure 7. Searching the Desulfovibrio vulgaris (Hildenborough) gene for cytochrome c3 in the Swissprot database with different indel error rates. Although the scores and the relative positions changed with different error rates, the same family members were recognized consistantly in top 10 of the returned sequence list.

slower than Smith-Waterman with the same computational complexity ($O(nm)$). To speed up the algorithm, we made some small modifications without affecting its performance significantly. When calculating a score matrix for a given frame, instead of considering all three matrix cells (i-1, j-1), (i-1, j) and (i, j-1) in other frames' matrices, we only consider the matrix cell (i-1, j-1) (see ***Figure 8***). The reason is that if there is a good partial match around the cell (i-1, j-1), it will be reflected in all three cells (i-1, j-1), (i-1, j) and (i, j-1), with slight differences (a insertion, a deletion or a substitution). So the result of considering just cell (i-1, j-1) should not differ significantly from considering all three cells. This reduces the running time by as much as 30% (to about one-fourth the speed of Smith-Waterman). In ***Figure 9***, the result of the modified algorithm is compared with that of the original algorithm.

4. Conclusion

This paper presented an algorithm for detecting and correcting frameshift errors that occur in protein encoding regions and can corrupt the recognition of important homologies. The algorithm appears to be robust and can recognize weakly related sequences in the presence of about 7% frameshift error and provides an optimal alignment. The algorithm provides a capability which can make single-pass or low-redundancy sequence data more informative, thereby reducing the necessity for high-redundancy sequencing for gene and protein characterization purposes. This work complements earlier work designed to recognize frameshifts in coding regions using pattern recognition [9], and both methods combined provide a powerful technology which can improve the efficiency and reduced the costs of genomic and cDNA sequencing.

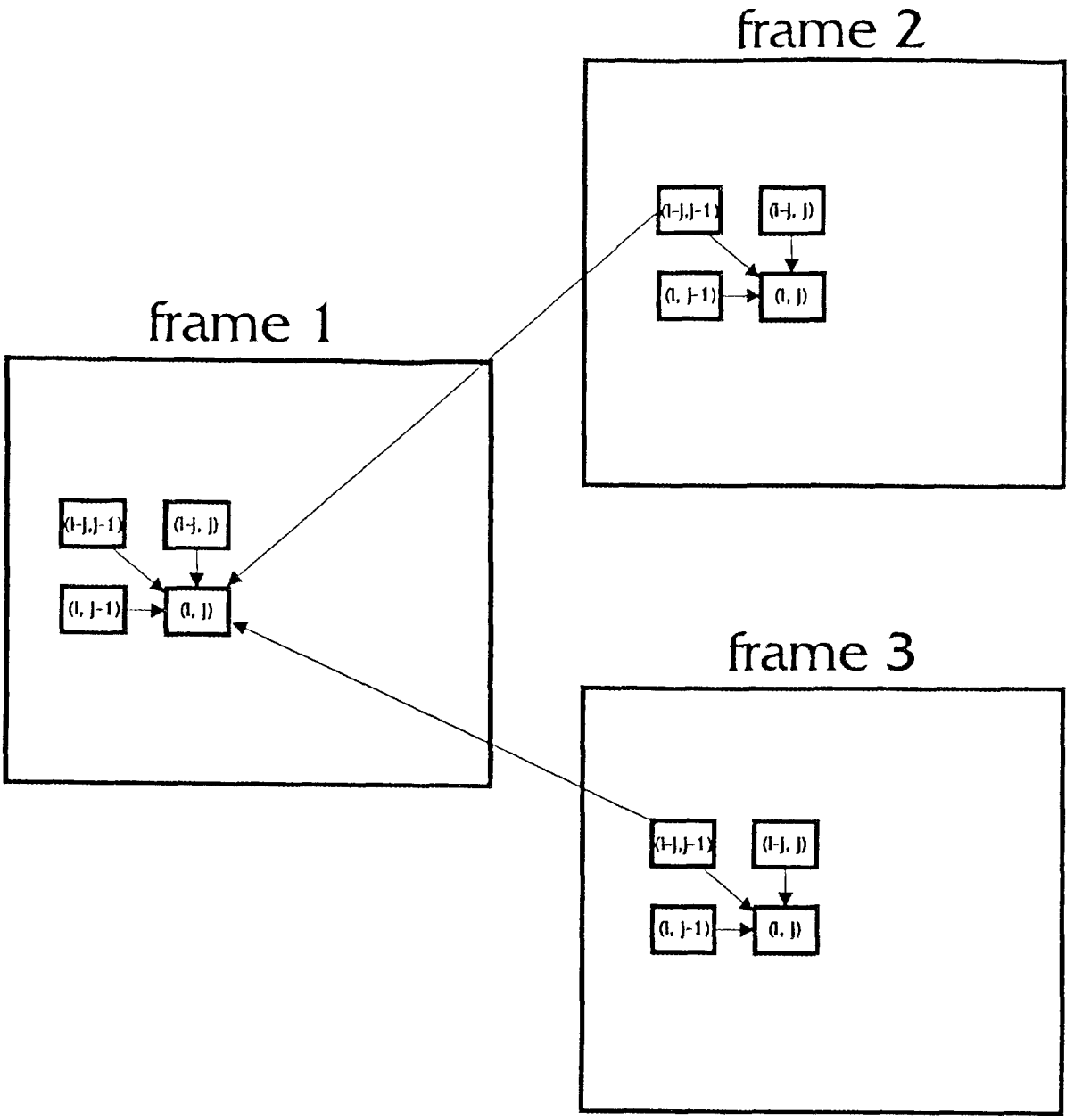
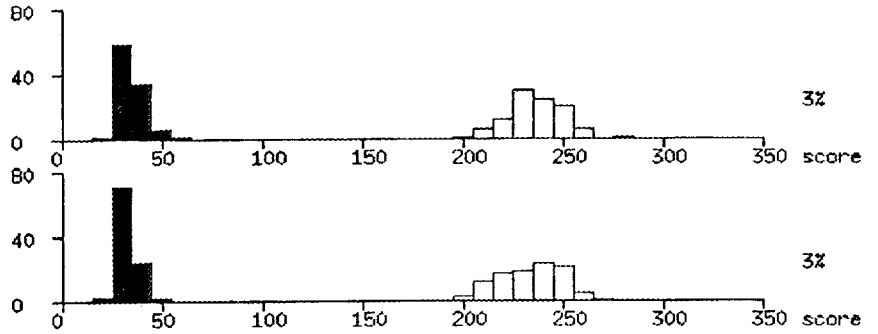


Figure 8. Schematic for the modified algorithm in which only one matrix cell (i-1, j-1) in other frames' matrices is considered in calculating matrix cell (i, j) for the present frame.

a. The effect of the modification on the performance with substitution errors: before (upper) and after (lower).



b. The effect of the modification on the performance with indel errors: before (upper) and after (lower).

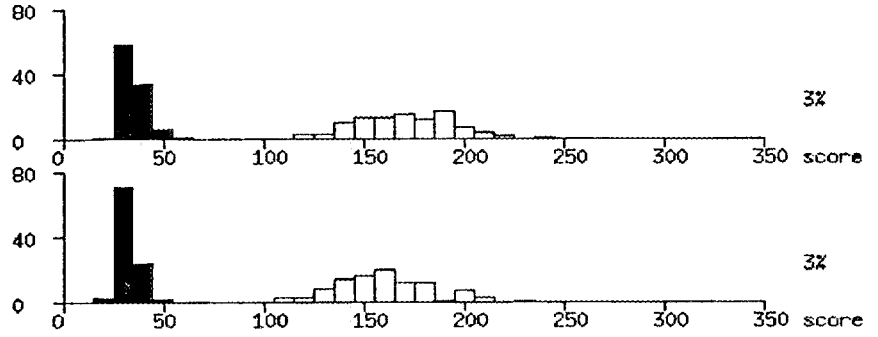


Figure 9. The modification of the algorithm has no significant effect on the performance of the algorithm for substitution (a) or frame shift errors (b).

5. Acknowledgements

This research was supported by the Office of Health and Environmental Research, United states Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

References

1. F. Altshcul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, 1990, Vol. 215, pp. 403-410.

2. Guan, Richard Mural, Sergey Petrov, and E. C. Uberbacher, "A Sensitive Sequence Comparison Server for DNA and PROTEINS," *Proc. Genome Sequencing and Analysis Conference V*, Hilton Head Island, South Carolina, 1993.

3. Gotoh, "An Improved Algorithm for Matching Biological Sequences," *Journal of Molecular Biology*, 1982, Vol. 162, pp. 705-708.

4. R. Pearson, and D. J. Lipman, "Improved Tools for Biological Sequence Comparison," *Proc. National Academy of Sciences, USA*, 1988, Vol. 85, pp. 2444-2448.

5. J. Posfai and R. J. Roberts, "Finding Errors in DNA Sequences," *Proc. National Academy of Sciences, USA*, 1992, Vol. 89, pp. 4698-4702.

6. R. F. Doolittle, "Searching Through Sequence Databases," in Doolittle, R. F. (ed.) *Methods in Enzymology*, 1990, Vol. 183, Academic Press, pp. 99-110.

7. D. J. States and D. Botstein, "Molecular Sequence Accuracy and the Analysis of Protein Coding Regions," *Proc. National Academy of Sciences, USA*, 1991, Vol. 88, pp. 5518-5522.

8. T. F. Smith, and M. Waterman, "Comparison of Biosequences," *Advances in Applied Mathematics*, 1981, Vol. 2, pp. 482-489.

9. Ying Xu, Richard J. Mural and Edward C. Uberbacher, "Correcting Sequencing Errors in DNA Coding Regions Using a Dynamic Programming Approach," *Cabios*, 1995 (in press).

10. A. Bairoch and B. Boeckmann, "The SWISS-PROT Protein Sequence Data Bank: Current Status," *Nucl. Acids Res.*, 1994, Vol. 22, pp. 3578-3580.

11. S. Henikoff and J. Henikoff, "Performance Evaluation of Amino Acid Substitution Matrices," *Proteins*, 1993, Vol. 17, 49-61.

INTERNAL DISTRIBUTION

1. M. Beckerman
2. J. R. Einstein
3. C. W. Glover
4. W. C. Grimmell
5-9. X. Guan
10. J. P. Jones
11. H. E. Knee
12-16. R. C. Mann
17. S. Matis
18. R. J. Mural
19. E. M. Oblow
20. C. E. Oliver
21. S. Petrov
22. N. S. V. Rao
23. D. E. Reichle
24. D. B. Reister
25. M. B. Shah
26. R. F. Sincovec
27-31. E. C. Uberbacher
32. R. C. Ward
33. X. Ying
34. EPMD Reports Office
35-36. Laboratory Records Department
37. Laboratory Records, ORNL-RC
38. Document Reference Library
39. Central Research Library
40. ORNL Patent Office

EXTERNAL DISTRIBUTION

41. Office of Assistant Manager for Energy Research and Development, Oak Ridge Operations, U.S. Department of Energy, P.O. Box 2008, Oak Ridge, TN 37831
42. Jim Decker, Director, Office of Energy Research, Dept. of Energy, Washington, DC 20585
43. David Smith, Health Effects & Life Sciences Research Division, Office of Health & Environmental Research, Dept. of Energy, Washington, DC 20585
44. B. J. Barnhart, Health Effects and Life Sciences Research Division, Office of Health and Environmental Research, Dept. of Energy, Washington, DC 20585
43. John Wooley, Health Effects and Life Sciences Research Division, Office of Health and Environmental Research, Dept. of Energy, Washington, DC 20585
44. Marvin Stodolsky, U.S. Department of Energy, Office of Health & Environmental Res., ER-72 GTN, Washington, DC 94720
45. Jay Snoddy, U.S. Department of Energy, Office of Health & Environmental Res., ER-72 GTN, Washington, DC 94720
46. Michelle Broido, U.S. Department of Energy, Office of Health & Environmental Res., ER-72 GTN, Washington, DC 94720
47. Robert Robbins, U.S. Department of Energy, Office of Health & Environmental Res., ER-72 GTN, Washington, DC 94720
48. Mark S. Boguski, National Institutes of Health NCBI/NLM, 8600 Rockville Pike, Bethesda, MD 20894
49. M. D. Zorn, Lawrence Berkeley Laboratory, MS 50B-3216, 1 Cyclotron Road, Berkeley, CA 94720
50. J. Fickett, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545

51. Christian Burks, Los Alamos National Laboratory, Theorectical Biol. & Biophysics Grp., T-10, MS K710, Los Alamos, NM 87545
52. A. Jamie Cuticchia, John Hopkins Hosp./Welch Med. Library, 1830 East Monument Street, Baltimore, MD 21205-2100
53. E. Branscomb, Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
54. A. Lapedes, Center for Human Genome Studies, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545
55. Radoje Drmanac, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439
56. Chris Fields, National Center for Genome Resources, 1800 Old Pecos Trail, Santa Fe, NM 87505

57-58. Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831

59. Raymond F. Gesteland, University of Utah, Howard Hughes Medical Institute, 6160 Eccles Genetics Building, Salt Lake City, UT 84112
60. Steven Henikoff, Fred Hutchison Cancer Research Ctr., HHMI-FHCRC, 1124 Columbia Street, Seattle, WA 98104
61. Tim Hunkapiller, University of Washington School of Medicine, Department of Molecular Biotechnology, FJ-20, Seattle, WA 98195
62. Leroy Hood, University of Washington School of Medicine, Department of Molecular Biotechnology, FJ-20, Seattle, WA 98195
63. Thomas G. Marr, Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, NY 11724
64. Sylvia J. Spengler, Lawrence Berkeley Laboratory, Human Genome Center, MS 1-213, 1 Cyclotron Road, Berkeley, CA 94720
65. Jude Shavlik, Computer Science, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706
66. Lawrence Hunter, Lister Hill Center, National Library of Medicine, Bethesda, MD 20892
67. Prof. Roger W. Brockett, Harvard University, Pierce Hall, 29 Oxford Street, Cambridge, MA 02138
68. Prof. Donald J. Dudziak, Dept. of Nuclear Engineering, 110B Burlington Engineering Labs, North Carolina State University, Raleigh, NC 27695-7909
69. Dr. James E. Leiss, Rt. 2, Box 142C, Broadway, VA 22815
70. Prof. Neville Moray, Dept. of Mechanical and Industrial Eng., University of Illinois, 1206 West Green Street, Urbana, IL 61801
71. Prof. Mary F. Wheeler, Department of Mathematics, Rice University, P.O. Box 1892, Houston, TX 77251
72. Peter Cartwright, Department of Human Genetics, University of Utah, Institute of Human Genetics, Room 2160, Bldg. 533, 10 North 2030 East Street, Salt Lake City, UT 84112
73. Norman Doggett, Genetics Group, Los Alamos National Laboratory, MS M880, LS-2, P.O. Box 1663, Los Alamos, NM 87545

74. Daniel Gusfield, Dept. of Computer Science, University of California, Davis, 3051 EU II, Davis, CA 95616
75. Anthony R. Kerlavage, The Institute for Genomic Research, 932 Clopper Road, Gaithersburg, MD 20878
76. Kenneth H. Fasman, Genome Data Base, Johns Hopkins University, Welch Medical Library, 2024 E. Monument Street, Baltimore, MD 21205
77. David Kingsbury, Genome Data Base, Johns Hopkins University, 2024 E. Monument Street, Baltimore, MD 21205
78. Chris Overton, Department of Genetics, University of Pennslyvania School of Medicine, CRB 475, 422 Curie Blvd., Ste. 1-200, Philadelphia, PA 19104-6145
79. David B. Searls, Department of Genetics, 475 CRB, Univ. of Pennsylvania School of Medicine, 422 Curie Boulevard, Philadelphia, PA 19104-6145
80. Randall Smith, Dept. of Molecular and Human Genetics, W.M. Keck Center for Computational Biology, Baylor College of Medicine, One Baylor Plaza, Room T921, Houston, TX 77030
81. Gary D. Stormo, Molecular, Cellular & Develop Biology, University of Colorado, Campus Box 347, Boulder, CO 80309-0347
82. David J. States, Biomedical Computer Lab, Washington University, Campus Box 8068, 700 South Euclid Avenue, St. Louis, MS 63110