ISE

ISE-TR-92-96

Similarity of DNA Sequences in DNA Databank Searching

by

Teruhisa Miyake, Sadaaki Miyamoto and Kazuhiko Nakayama

February 29, 1992

INSTITUTE
OF
INFORMATION SCIENCES AND ELECTRONICS
UNIVERSITY OF TSUKUBA

Similarity of DNA Sequences in DNA Databank Searching

Teruhisa Miyake, Sadaaki Miyamoto*, and Kazuhiko Nakayama

Institute of Information Sciences and Electronics
University of Tsukuba, Ibaraki 305, Japan

*Faculty of Engineering, University of Tokushima
Tokushima 777, Japan

ABSTRACT

The homology searching is an effective technic when one want to analyze the similarity of DNA sequences but it cannot be applied to a DNA databank searching directly. In this paper, we discussed on a similarity of DNA sequences by which the DNA databank can be retrieved fuzzyly and we proposed two type of new similarity definitions and reported an experimental databank in which the similarity proposed here are implemented.

Keywords: Matching function, homology searching, similarity,
DNA databank, Jaccard coefficient, fuzzy retrieval

1. Introduction.

As the number of DNA sequences has grown into the millions, it is clear that similarity searching of DNA databank becomes very important in molecular biology. For the purpose of detailed analysis of two sequences, there are many methods of homology searching. For the global alignment (compares complete sequences), the methods of Needleman and Wunsch[1], of Sankoff[2], of Fitch and Smith[3] and of Sellers[4] are well known. For the local alignment (compares subsequences of two sequences), the methods of Brutlag et. al.[5], of Smith and Waterman[6] and of Goad and Kanehisa[7] are known widely. But those famous methods can not be apply to a databank searching directly. For example, the method of Needleman and Wunsch needs three parameters named "match value", "mismatch value" and "gap penalty" in execution, and the result value that indicates the similarity of the sequences changes if those parameter value changed. In other word, we must select suitable values for those parameters to get a reasonable result. But the suitable values of searching changes if sequence pair is changed. By this property, this method is difficult to apply a databank searching that has to compare with various sequences in a same standard. In this point, all of the methods mentioned above have similar difficulties. Consequently we need a new method that is suitable for a databank searching. This paper proposes a new concept for the similarity in a databank searching and discusses its characteristics to implement real DNA databank.

2. Homology searching

For the example of the method of homology searching, we consider the method of Needleman and Wunsch named maximum similarity search(MSS) method, it is mentioned above. When a sequence A and a sequence B are compared by this method, we make a L × N sized matrix M named "matrix of match" (Let L and N be a length of A and B respectively). Its element m_{ij} is set to a value of parameter "match" if i_{th} position of A is equal to j_{th} position of B and set to the value of "mismatch" if i_{th} position of A is not equal to j_{th} position of B.(Figure 1). Then we make a matrix S named "matrix of score", its size is the same above. Its element s_{ij} is set a value or a "score" described below.

First, the first row elements and the first column elements are set to the same values of corresponding elements of M. The scores of other elements are determined by the score of its three neibouring elements. The score of s_{ij} is set to the maximum value of the three scores below:

1) the score of $s_{i-1j-1} + M_{ij}$
2) the score of $s_{ij-1} + M_{ij}$ - "gap penalty"
3) the score of $s_{i-1j} + M_{ij}$ - "gap penalty"

This process is repeated to the all elemets in the matrix S (Figure 2). The highest score of the Lth row elements and the Nth column elements named "maximum score" shows a similarity value of the sequence pair and from the "maximum score" element there are several pathes to the other side row or column. Those pathes show the common part of two sequences by mean of this score and are named "optimal alignment"s.(Figure 3) An Optimal alignment of DNA sequences is shown in Figure 4. This method uses a dynamic programming thechnic and is efficient to execute on computer systems. But, as explained above, the result is changed by the selection of the parameter values and the suitable parameter values are not determined easily. Besides, the meaning of the parameters is not clear for users . Futhermore, suitable parameter values of one comparing cannot be suitable to the other comparing. This property makes difficult to apply this method to a databank searching.

3. Required properties of the similarity for a databank searching

The DNA sequences are very long and this means that it costs much resources to culculate the similarity. For that reason the efficiency to culculate the similarity is very important. The number of a DNA databank is enormous so a method to culculate similarity that need some parameter arrangement at every comparing sequences cannot be used for retrieval. The retrieval system must execute automatically. A user of DNA databank may not be a specialist of molecular biology so the retrival system of DNA databank has not to expect adding informations from a user. In other word, a user need not to arrange any parameter at a searching. Consequently, to select a method to culculate similarity, we must consider,

1)efficiency of culculation
2)execution automatically

3)necessity for parameters that must be arranged at a searching

The similarity defined in the MSS method is good at the first requirement but it does not satisfy the second and the third requrement. Therefore we must consider an other similarity definition suitable for a databank searching.

4. Form estimate method

Instead of using the maximum score of the MSS method, we try a variation of the method. When two sequences are similar, a form of its optimal alignment is strait and long and has few gaps and when two sequences are very different, that has a winding and complexed shape and has many gaps. In other word, the optimal alignment shows how those sequences are similar or not. If we can estimate the form of the optimal alignment, then we can estimate the similarity of the two sequences. The estimate value of the form of the optimal alignment can be used as a similarity value. Certainly, this estimation is justified only when those two sequences are almost similar and a form of an optimal alignment of very different sequence pair cannot be estimated in this way. But it is not important that the estimation is not suitable for every comparing. In the databank searching we need only the most similar sequences. Then, we define three grades to estimate a form of optimal alignment as below:

1)grade of winding : W

$$W = \sum_{i=1}^{t} w_i$$

wi : distance of an element of the optimal alignment and
diagonal line of the sore matrix
t : total number of elemt of optimal alignment

2)grade of length : L

$$L = \frac{\text{length of optimal alignment}}{\text{length of diagonal}}$$

3)grade of match : M

$$M = \frac{\text{number of matched elements in optimal alignment}}{\text{total number of elements in optimal alignment}}$$

$(0 \leq W, L, M \leq 1)$

If compared two sequences are very similar, the value of each grade becomes near 1.0. We consider those grades as fuzzy membership functions of the similarity of the sequences and a membership function of the similarity can be made from a combination of tsose membership functions. Then we define a membership function of the similarity s_p as a intersection of the three fuzzy sets:

$$s_p = W \cap L \cap M$$

The value of s_p also depends on the parameters, "match", "unmatch", "gap penalty". This means that the membership function s_p is defined in the space of those three parameters. Therefore we cannot apply s_p to a databank searching directly. Lastly we define overall similarity s which does not depend on any parameter and can be applied to a databank searching.(Figure 5)

$$s = \max_p (s_p)$$

5. matching functions

A DNA sequence is a string of nucleotides of 4 types, A, T, C, G. From the point of information science view, it can be regarded as a long character string which its alphabet has only four characters, A, T, C, G. Then the similarity of the two DNA sequence can be defined a similarity of character strings. A matching function of character strings is used in information retrieval to match index terms and keywords. Usually it is a crisp type function so its value is 0 or 1 and it cannot be used to culculate similarity value. But matching functions used in fuzzy information retrieval can be applied to this purpose. The difference between a normal keyword and a DNA sequence is the length. The length of DNA sequences are much longer than normal keywords. Its length can be reached to thousands or even millions elements. Because of execution efficiency, we select functions that needs no matrix culculations but based on set thoeretical operations. We regard a string as a set of substrings included itself and a matching function is composed of its operations. This type of matching functions has been studied in information science and well-known functions are shown in Table 1.

Table 1 Matching functions

Jaccard Coefficient	$s_{jc} = \dfrac{\mid A \cap B \mid}{\mid A \cup B \mid}$
Dice Coefficient	$s_d = \dfrac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid}$
Cosine Coefficient	$s_{cos} = \dfrac{\mid A \cap B \mid}{\mid A \mid^{1/2} \cdot \mid B \mid^{1/2}}$
Overlap Coefficient	$s_o = \dfrac{\mid A \cap B \mid}{\min [\, \mid A \mid, \mid B \mid \,]}$

(A, B : set, |A| : cardinality of A)

The charactaristics of those functions applied to the set of substrings are shown in Fugure 6 and Figure 7 by a short string matching example. We select Jaccard coefficient as a matching function, because it is more sensitive for the length of matched substrings than others. we consider it is effective to distinguish a similar string from different strings.

Then the matching function s is defined as below:,

$$s = \frac{\mid \mathrm{sub}\,(\mathrm{str}_1) \cap \mathrm{sub}\,(\mathrm{str}_2) \mid}{\mid \mathrm{sub}\,(\mathrm{str}_1) \cup \mathrm{sub}\,(\mathrm{str}_2) \mid}$$

str_1 and str_2 : the DNA sequences to be compared.

sub(A) : the set of substrings of A.

Let L_1 and L_2 be the length of str_1 and str_2 respectively and be assumed $L_1 \leq L_2$ and $L \cong L_1 \cong L_2$. Then the number of calculation C is

$$C = \sum_{i=1}^{L_1} (L_1 - i + 1) (L_2 - i + 1) \qquad \approx o\,(L^3/ 3)$$

As the string is very long, C becomes a huge number, but we can culculate it more efficiently by arranging this method to sort substrings before searching.

6. Experimental databank

For the evaluation of the two similarity definition, we had implemented these to an experimental databank system which contains 26 DNA sequences. Those sequences are selected from EMBL (European Molecular Biology Laboratory) database. This system is working on FACOM M1800 computer. An example of databank searching using the Form estimation method is shown in figure 8 and that using a matching function is shown in Figure 9.

7. Conclusion

The MSS method is efficient to a typical homology searching, but it cannot be applied for a DNA databank searching. In this paper, we proposed two methods to culculate a similarity of DNA sequences. One is the method of the form estimation of optical alignment and the other is the method using a matching function. Both methods can be applied to a databank searching and the method of the form estimation is more efficient than using a matching function now. But each method has different characteristics, for exmple the former costs little CPU time and large memories and the latter costs long CPU time and little memories. One can select the suitable method by one's own system conditions. We discussed here on the similarity of whole sequences i.e. the global alignment search, but the methods proposed here are also effective to realize the local alignment searching of a databank.

references

1. Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two protains, J. Med. Biol., 48(1970), 443-453

2. Sankoff, D. Matching sequences under Deletion/Insertion constraints, Proc. natn. Acad. Sci. U.S.A., 69(1972), 4-6

3. Fitch, W. and Smith, T. Optimal Sequence Alignment, Proc. natn. Acad. Sci. U.S.A., 80(1983), 1382-1386

4. Sellers, P. H. On the theory and computation of evolutionary distances, SIAM J. Appl. Math., 26(1974), 787-793

5. Brutlag, D. L., Clayton, J., Friedland, P. and Kedes, L. H. SEQ: a nucleotide sequence analysis and recombination system, Nucl. Acids Res., 10(1982), 279-294

6. Smith, T. F. and Waterman, M. S. Idntification of common molecular subsequences, J. molec. Biol., 147(1981), 195-197

7. Goad, W. B. and Kanehisa, M. I. Pattern recognition in nucleic acid sequences I. A general method for finding local homologies and symmetry, Nucl. Acids Res., 10(1982), 247-263

8. Miyake, T., Miyamoto, S. and Nakayama, K. Application of a fuzzy matching function to DNA sequence analysis, 6th Fuzzy System Symposium(1990), 307-310

9. Miyake, T., Takahashi, T. Homology search of DNA sequence data based on pattern evaluation of optimal alignments,7th Fuzzy System Symposium(1991), 367-370

Sequence A

Sequence B	G	T	T	C	A	C
G	2	-1	-1	-1	-1	-1
G	2	-1	-1	-1	-1	-1
T	-1	2	2	-1	-1	-1
C	-1	-1	-1	2	-1	2
C	-1	-1	-1	2	-1	2

(Match=2, Mismatch=-1, Gap penalty=-1)

Figure 1 matrix of match

Sequence A

Sequence B	G	T	T	C	A	C
G	2	-1	-1	-1	-1	-1
G	2	1	-1	-2	-2	-2
T	-1	4	5	3	1	-1
C	-1	2	3	7	5	6
C	-1	0	1	8	6	7

Figure 2 matrix of score

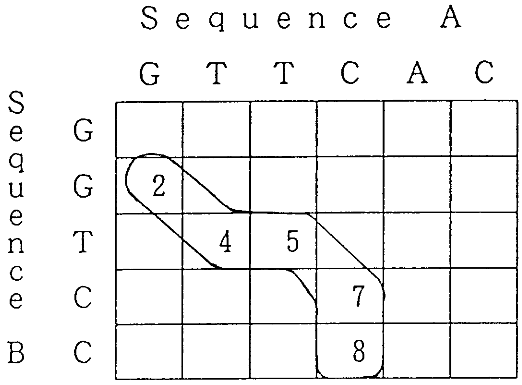
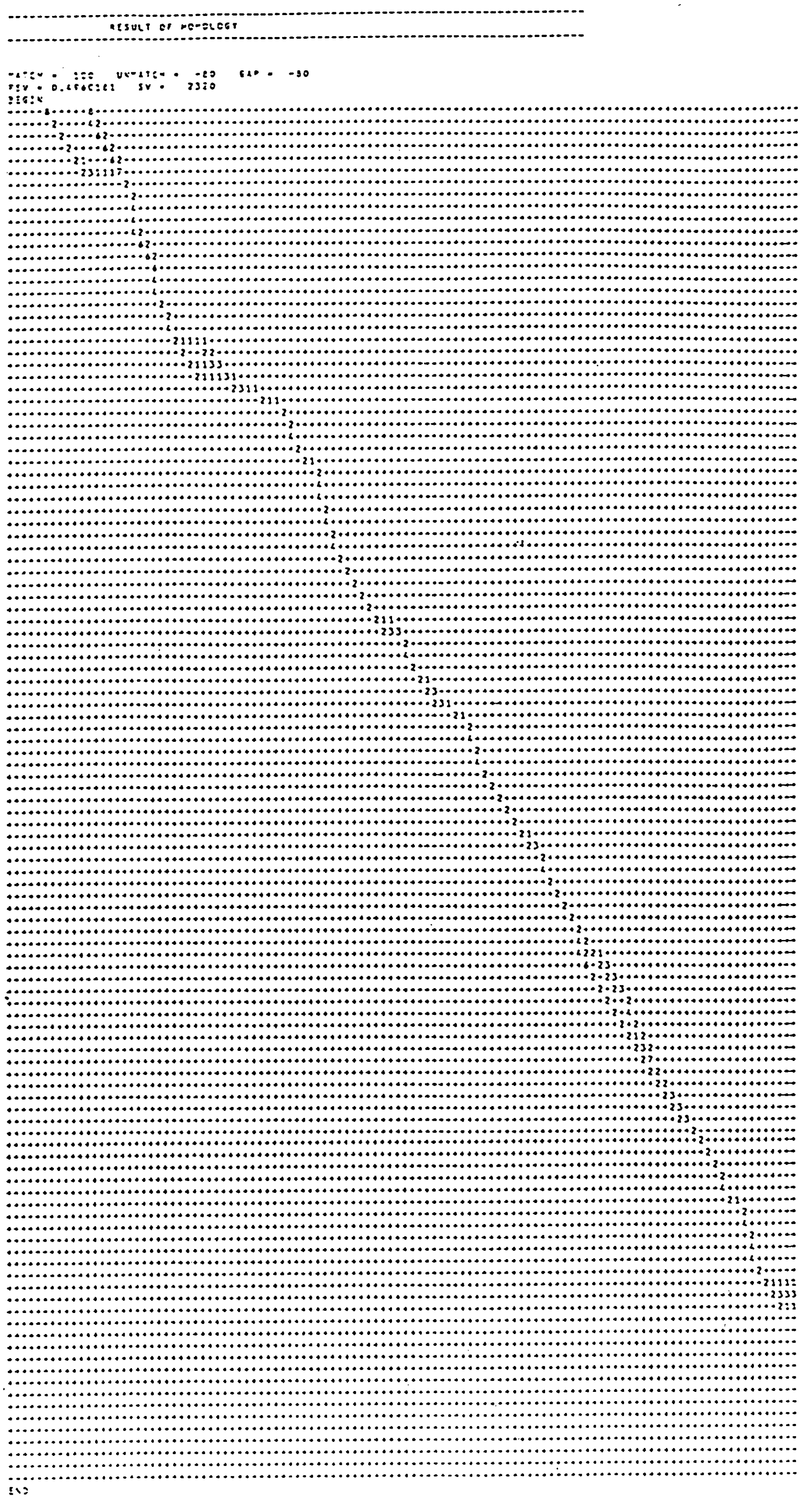


Figure 3 Optimal alignment



```
RESULT OF HOMOLOGY

MATCH =  100   UNMATCH =  -20   GAP =  -50
FSV = 0.4546C161   SV =  2320
BEGIN
```

```
END
```

Figure 4 optimal alignment of PAMVM1 and GGCOL2

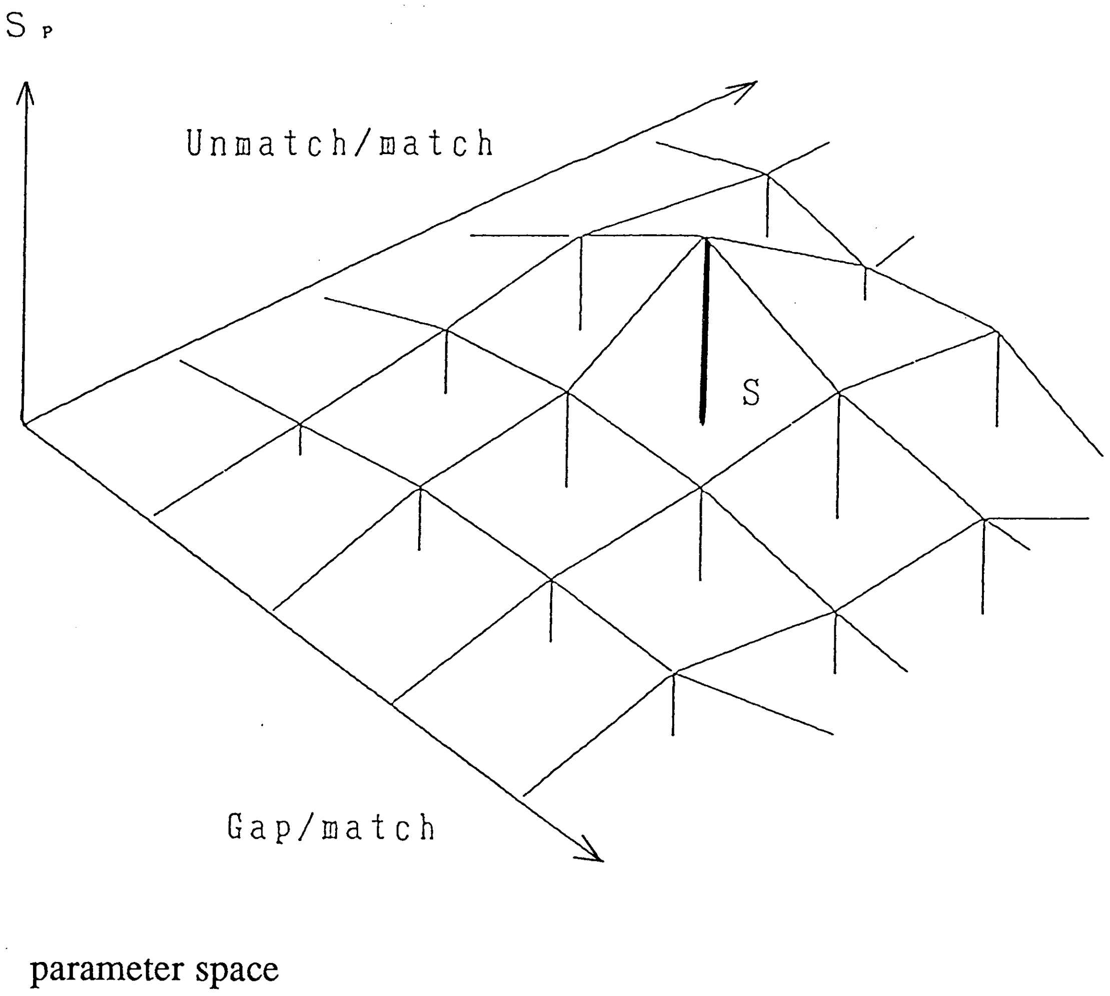


Figure 5 Similarity s

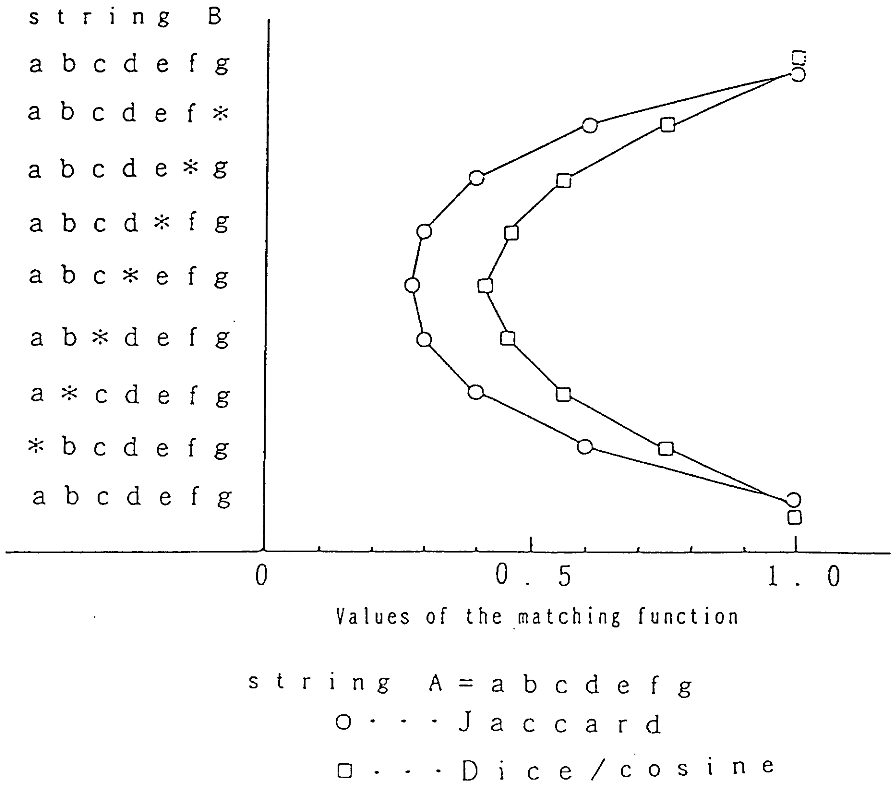


Figure 6 Value of matching functions(1)

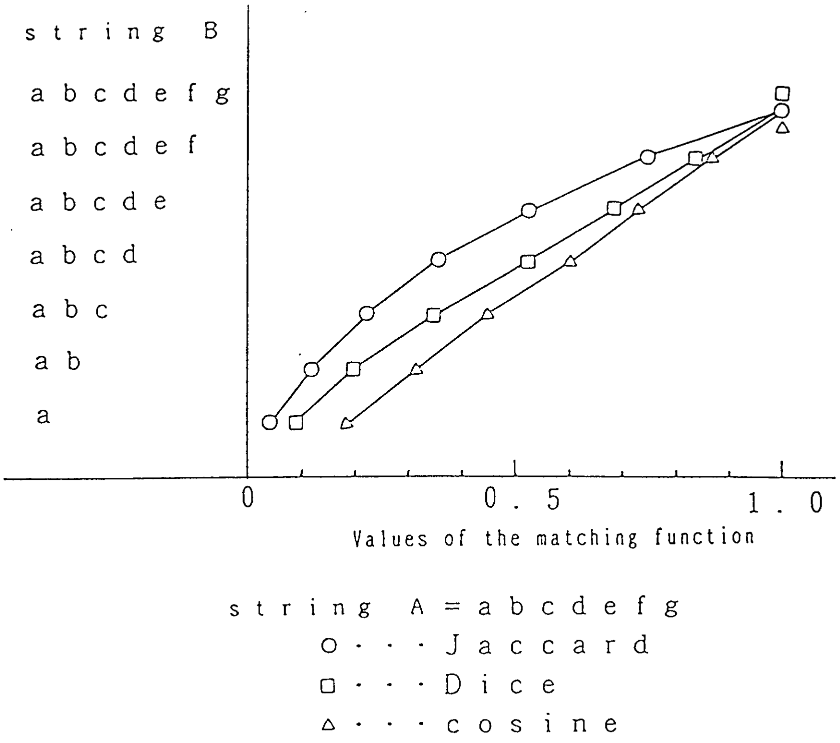
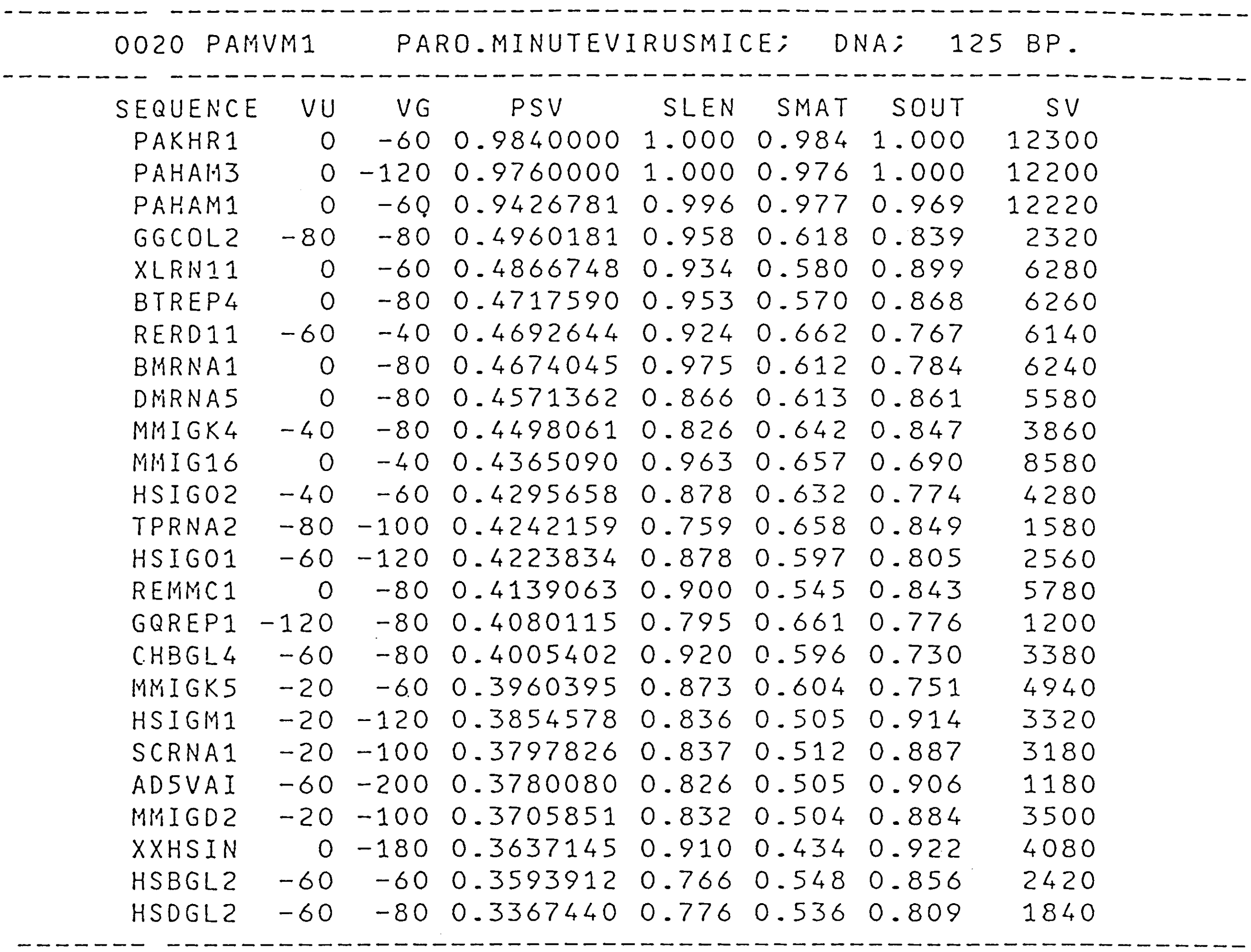
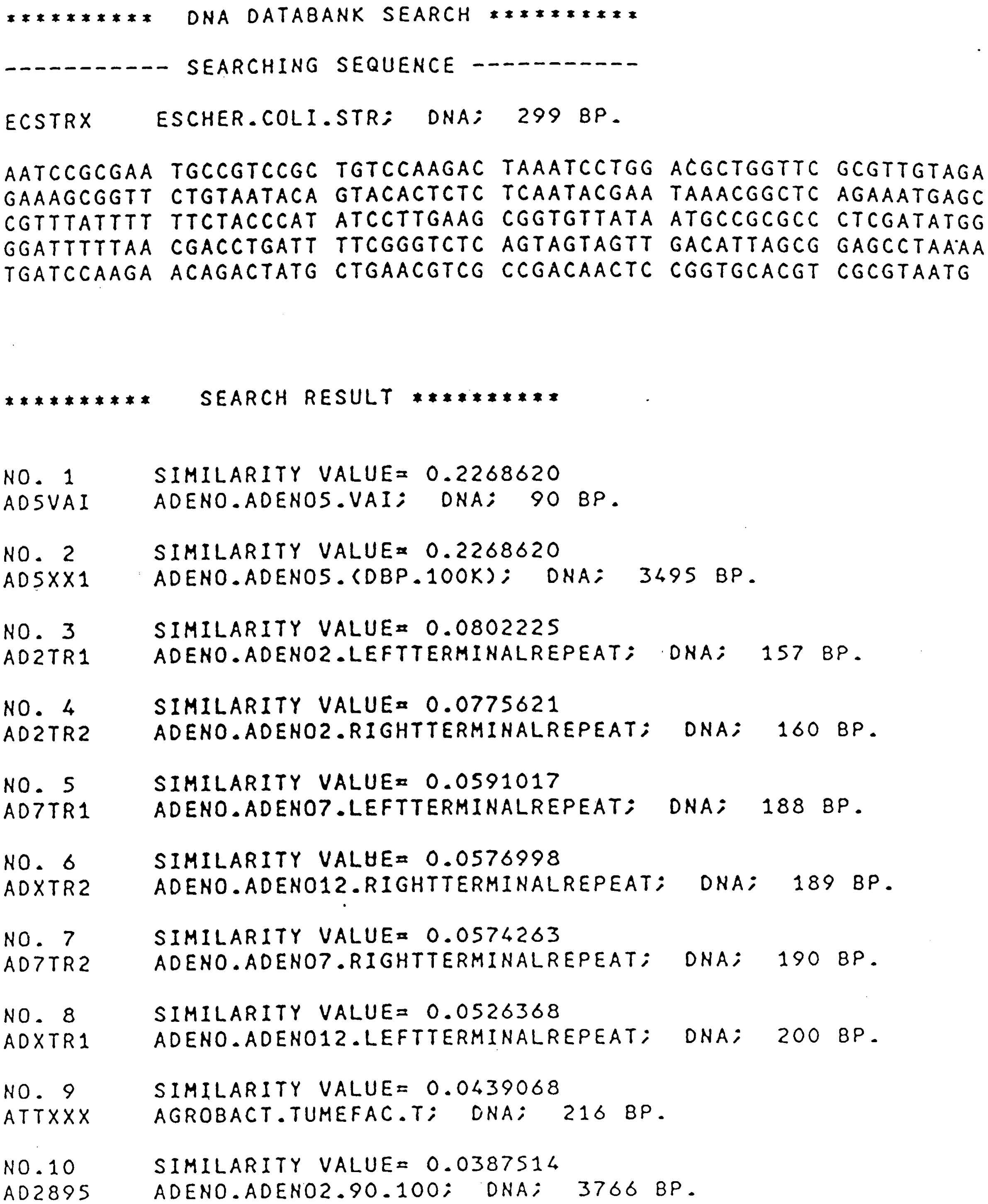


Figure 7 Value of matching functions(2)



```
--------  ----------------------------------------------------------
     0020 PAMVM1     PARO.MINUTEVIRUSMICE;  DNA;   125 BP.
--------  ----------------------------------------------------------
     SEQUENCE   VU    VG     PSV      SLEN   SMAT   SOUT     SV
      PAKHR1     0   -60 0.9840000 1.000 0.984 1.000  12300
      PAHAM3     0  -120 0.9760000 1.000 0.976 1.000  12200
      PAHAM1     0   -60 0.9426781 0.996 0.977 0.969  12220
      GGCOL2   -80   -80 0.4960181 0.958 0.618 0.839   2320
      XLRN11     0   -60 0.4866748 0.934 0.580 0.899   6280
      BTREP4     0   -80 0.4717590 0.953 0.570 0.868   6260
      RERD11   -60   -40 0.4692644 0.924 0.662 0.767   6140
      BMRNA1     0   -80 0.4674045 0.975 0.612 0.784   6240
      DMRNA5     0   -80 0.4571362 0.866 0.613 0.861   5580
      MMIGK4   -40   -80 0.4498061 0.826 0.642 0.847   3860
      MMIG16     0   -40 0.4365090 0.963 0.657 0.690   8580
      HSIG02   -40   -60 0.4295658 0.878 0.632 0.774   4280
      TPRNA2   -80  -100 0.4242159 0.759 0.658 0.849   1580
      HSIG01   -60  -120 0.4223834 0.878 0.597 0.805   2560
      REMMC1     0   -80 0.4139063 0.900 0.545 0.843   5780
      GQREP1  -120   -80 0.4080115 0.795 0.661 0.776   1200
      CHBGL4   -60   -80 0.4005402 0.920 0.596 0.730   3380
      MMIGK5   -20   -60 0.3960395 0.873 0.604 0.751   4940
      HSIGM1   -20  -120 0.3854578 0.836 0.505 0.914   3320
      SCRNA1   -20  -100 0.3797826 0.837 0.512 0.887   3180
      AD5VAI   -60  -200 0.3780080 0.826 0.505 0.906   1180
      MMIGD2   -20  -100 0.3705851 0.832 0.504 0.884   3500
      XXHSIN     0  -180 0.3637145 0.910 0.434 0.922   4080
      HSBGL2   -60   -60 0.3593912 0.766 0.548 0.856   2420
      HSDGL2   -60   -80 0.3367440 0.776 0.536 0.809   1840
-------  -----------------------------------------------------------
```

Figure 8 An example of the form estimation method



```
**********  DNA DATABANK SEARCH **********

----------- SEARCHING SEQUENCE -----------

ECSTRX     ESCHER.COLI.STR;  DNA;  299 BP.

AATCCGCGAA TGCCGTCCGC TGTCCAAGAC TAAATCCTGG ACGCTGGTTC GCGTTGTAGA
GAAAGCGGTT CTGTAATACA GTACACTCTC TCAATACGAA TAAACGGCTC AGAAATGAGC
CGTTTATTTT TTCTACCCAT ATCCTTGAAG CGGTGTTATA ATGCCGCGCC CTCGATATGG
GGATTTTTAA CGACCTGATT TTCGGGTCTC AGTAGTAGTT GACATTAGCG GAGCCTAAAA
TGATCCAAGA ACAGACTATG CTGAACGTCG CCGACAACTC CGGTGCACGT CGCGTAATG



**********   SEARCH RESULT **********


NO. 1      SIMILARITY VALUE= 0.2268620
AD5VAI     ADENO.ADENO5.VAI;  DNA;  90 BP.

NO. 2      SIMILARITY VALUE= 0.2268620
AD5XX1     ADENO.ADENO5.(DBP.100K);  DNA;  3495 BP.

NO. 3      SIMILARITY VALUE= 0.0802225
AD2TR1     ADENO.ADENO2.LEFTTERMINALREPEAT;  DNA;  157 BP.

NO. 4      SIMILARITY VALUE= 0.0775621
AD2TR2     ADENO.ADENO2.RIGHTTERMINALREPEAT;  DNA;  160 BP.

NO. 5      SIMILARITY VALUE= 0.0591017
AD7TR1     ADENO.ADENO7.LEFTTERMINALREPEAT;  DNA;  188 BP.

NO. 6      SIMILARITY VALUE= 0.0576998
ADXTR2     ADENO.ADENO12.RIGHTTERMINALREPEAT;  DNA;  189 BP.

NO. 7      SIMILARITY VALUE= 0.0574263
AD7TR2     ADENO.ADENO7.RIGHTTERMINALREPEAT;  DNA;  190 BP.

NO. 8      SIMILARITY VALUE= 0.0526368
ADXTR1     ADENO.ADENO12.LEFTTERMINALREPEAT;  DNA;  200 BP.

NO. 9      SIMILARITY VALUE= 0.0439068
ATTXXX     AGROBACT.TUMEFAC.T;  DNA;  216 BP.

NO.10      SIMILARITY VALUE= 0.0387514
AD2895     ADENO.ADENO2.90.100;  DNA;  3766 BP.
```

Figure 9 An example of using a matching function

INSTITUTE OF INFORMATION SCIENCES AND ELECTRONICS
UNIVERSITY OF TSUKUBA
TSUKUBA-SHI, IBARAKI 305 JAPAN

REPORT DOCUMENTATION PAGE	REPORT NUMBER ISE-TR-92-96

TITLE

Similarity of DNA Sequences in DNA Databank Searching

AUTHOR(S)

Teruhisa Miyake, Sadaaki Miyamoto*, and Kazuhiko Nakayama

Institute of Information Sciences and Electronics
University of Tsukuba, Ibaraki 305, Japan
*Faculty of Engineering, University of Tokushima
Tokushima 777, Japan

REPORT DATE 1992-2-29	NUMBER OF PAGES 14
MAIN CATEGORY Information Search and Retrieval	CR CATEGORIES

KEY WORDS
Matching function, homology searching, similarity, DNA databank, Jaccard coefficient, fuzzy retrieval

ABSTRACT

The homology searching is an effective technic when one want to analyze the similarity of DNA sequences but it cannot be applied to a DNA databank searching directly. In this paper, we discussed on a similarity of DNA sequences by which the DNA databank can be retrieved approximately or fuzzyly and we proposed two type of new similarity definitions and reported an experimental databank in which the similarity proposed here are implemented.

SUPPLEMENTARY NOTES