TR-728

MASCOT: Multiple Alignment System for Protein Sequences based on 3-way Dynamic Programming

by

M. Hirosawa, M. Hoshida, M. Ishikawa
& T. Toya

January, 1992

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191 ~ 5
Telex ICOT J32964

Institute for New Generation Computer Technology

MASCOT : Multiple Alignment System for Protein Sequences based on 3-way Dynamic Programming

Makoto Hirosawa, Masaki Hoshida, Masato Ishikawa, Tomoyuki Toya
(Institute for New Generation Computer Technology (ICOT),
1-4-28 Mita,Minato-ku, Tokyo 108, Japan)

Key Words : Multiple Alignment, 3-way DP, Simulated Anealing, Clustering

Abstract
A multiple alignment methodology that can produce high quality alignment is extremely important for predicting the structure of unknown proteins. Nearly all the methodologies developed so far have employed 2-way alignment only. Though these methods are fast, the alignments they produce lose reliability as the similarity of sequences reduces.
We developed the MASCOT multiple alignment system. MASCOT can sustain the reliability of alignment even when the similarity of sequences is low. MASCOT achieves high-quality alignment by employing 3-way alignment in addition to 2-way alignment. The resultant alignments are refined by Simulated Annealing to futher quality . We also use a cluster analysis of sequences to produce highly reliable alignments.

1 Introduction

The similarity analysis of protein sequences by the use of multiple alignment is an important technique for predicting the function and structure of proteins for drawing phylogenetic trees of creatures. Until recently, multiple alignment was produced by biologists. However with the increasing rate of determination of protein sequences, computer assitance in multiple alignment is becoming indispensable.

It is well-known that once a similarity index between amino acids is given, the multiple alignment problem can be solved theoretically by Dynamic Programming (DP) (Neeydleman and Wunsch,1970). N-way DP can align n sequences simultaneously and can derive the optimal alignment of these sequences.

One problem with DP is the incredible computational time it requires. N way DP takes computational time in the order of the n-th power of the sequence length . To keep this expansion of computational time manageable, nearly all multiple alignment systems developed so far employ 2-way DP as a base and combine the results of 2-way DP to produce multiple alignment (Barton, 1990).

This class of alignment methods is good because of the small computational time required, but this is not sufficient to produce an alignment of sequences when similarity is low.

To produce multiple alignments of high-quality with small increases in computational time, we sought another class of alignment scheme : a multiple alignment system based on 3-way DP. We named it MASCOT (Multiple Alignment System developed by iCOT). The characteristics of MASCOT are (1)Production of *initial alignment* with *core alignment* by the use of 3-way DP, (2)*Refinement* of initial alignment by the use of *Simulated Annealing*, (3)Cluster analysis of sequences using triplet of sequences as a unit.

Each module of MASCOT is described by the KL1 (Nakajima *et al.*, 1989) parallel logic language and is executed on the Multi PSI (Nakajima *et al.*, 1989) parallel inference machine. Though MASCOT requires more computation than conventional alignment systems due to the use of 3-way DP, parallel execution by the parallel inference machine (Ishikawa *et al.*,1991a) can reduce the total time.

In this paper, an overview of MASCOT is introduced in the 2nd chapter. Then three features of MASCOT are explained in the 3rd and the 4th and the 5th chapters. Finally, in the 6th chapter, an example of MASCOT application to retro-viral sequences are shown.

2 Multiple Alignment System MASCOT

MASCOT is composed of four kinds of modules, namely, (1)"Cluster Analysis", (2)"Intra-cluster Alignment",(3)"Inter-cluster Alignment" and (4)"Evaluation of Alignment"(upper part of Fig.1). We will explain briefly the structure of MASCOT as a whole.

Multiple alignment is the problem of aligning similar residues of many sequences in the same column. Usually similarity of two sequences to be focused among the sequences are different. Some pairs of sequences have high similarity while others have low similarity. When sequences of low similarity are aligned forcibly, the resulting alignment is often unreliable. The general strategy taken so far has been to group sequences according to similarity, and to align sequences in the order of their similarity (Barton, 1990).

MASCOT also uses this strategy. Firstly, in "Cluster Analysis", sequences are grouped into plural clusters according to their similarity. Secondly, in "Intra-cluster Alignment", sequences belonging to the same cluster are aligned. Thirdly, in "Inter-cluster Alignment", the resultant alignments of individual clusters are combined to produce alignment of the whole. Finally, in "Evaluation of Alignment", the resultant alignments are evaluated. Both the order of intra-cluster alignment and that of inter-cluster alignment are decided in "Cluster Analysis"(See Chapter 5).

"Intra-cluster Alignment"(lower part of Fig.1) uses strategy of *two phase alignment* (Hirosawa *et al.*, 1991). In the first phase initial alignment is produced by considering a few strongly-similar sequences. In the second phase, each initial alignment is refined by considering all sequences.

It is not practical to align sequences refinedly due to the problem of computational time. Therefore some of the multiple alignment systems firstly developed so far align roughly, and then refine this rough initial alignment manually or by the use of a computer (Barton, 1990). Unless the quality of the initial alignment is above a certain level, the refinement phase doesn't work effectively and this strategy fails.

To make two phase alignment effective, MASCOT has adopted an initial alignment scheme based on 3-way DP (See Chapter 3). And, in the refinement phase, we adopted Simulated Annealing that is an effective stochastic optimization method(See Chapter 4).

"Inter-cluster Alignment"(upper part of Fig. 1) combines intra-cluster alignment to align all sequences given to MASCOT. In "Inter-cluster Alignment", DP between groups of sequences is executed iteratively.

If there are plural alignments when "Inter-cluster Alignment" has finished, "Evaluation of Alignment" selects the most reliable alignments and show them to the users. There are two cases when plural multiple alignments are produced. One case is when some stage of MASCOT produces equivalently evaluated alignments. The other case is when we adopt a strategy to retain plural candidates of these multiple alignments which are evaluated highly. The latter strategy is taken because the valuation index based on Dayhoff Score(Dayhoff *et al.*, 1978) adpopted by MASCOT is not absolute. In other words, it is necessary to consult biological knowledge to evaluate alignments properly in this case.

"Evaluation of Alignment" evaluates alignments comprehensively using biological knowledge. For example, the length and the number of conserved regions in alignment and whether the conserved regions contain biological meaningful motifs are evaluated. MAS-

COT is experimentally equipped with a motif dictionary and already discoverd information can be reflected in the multiple alignment.

3 Generation of Initial Alignment based on 3-way DP

To produce an initial alignment, MASCOT firstly makes a core alignment composed of a few sequences based on 3-way DP and then attaches other sequences one by one. The sequences selected as core sequences must be the most similar sequences. The selection of core sequences and the order in which sequences are attached is decided by "Cluster Analysis"(See Chapter 5).

The quality of alignment using a core alignment depends on the reliability of the core alignment. The number of core sequences should not be small. Aligning these sequences using DP simultaneously is one ways to correctly make core alignment. Due to the restrictions of computational time, however, pratically implementable DP is up to 3-way DP. In consideration of this, we have decided to sellect four core sequences and to combine two alignments produced by 3-way DP to make core alignment.

3-way DP merging method is the name of the method that combine two alignments produced by 3-way DP (Ishikawa *et al.*, 1991a; Hirosawa *et al.*, 1991). We will explain this method using a small size example(Fig.2). Fig.2 shows the steps from two multiple alignments of three sequences by 3-way DP (Step 0) to a core alignment of four sequences(Step 3).

In Step 1, alignment of HTLV-2 and BLV included in the alignment of {HTLV-1, HTLV-2, BLV} and that included in the alignment of {HTLV-2, BLV, HIV} are compared and consistent regions are searched for. In the figure, *the consistent regions* are marked '1', and *the inconsistent regions* are marked '0'. Here '.' is a gap introduced to align consistent regions. In inconsistent region, alignment of the part of the sequences of HTLV-2 and BLV is [SLL,VHL] in the upper alignment, but is [-SLL,VHL-] in the lower alignment. It indicates contradiction.

In Step 2, four sequences {HTLV-1, HTLV-2, BLV, HIV} are aligned by a simple procedure in consistent regions and by a complex procedure in inconsistent regions. In the consistent region of on left side of the alignment [LLQAI,VLQAI,LLEAI,FL--L] is constructed by simply combining [LLQAI,VLQAI,LLEAI] (from the upper alignment) and [VLQAI,LLEAI,FL--L] (from the lower alignment). In the right hand consistent region, the same procedure is executed.

In the inconsistent region in the middle where the upper alignment is [AHL,SLL,VHL] and the lower alignment is [-SLL,VHL-,-KLA], four kinds of alignment are computed by DP between two groups of sequences and the alignment of the highest evaluation is selected as the alignment of the inconsistent region.

(i) [AHL,SLL,VHL] (from the upper alignment) and [-KLA] (from the lower alignment)

(ii) [AHL,SLL] (from the upper alignment) and [VHL-,-KLA] (from the lower alignment)

(iii) [AHL,VHL] (from the upper alignment) and [-SLL,-KLA] (from the lower alignment)

(iv) [AHL] (from the upper alignment) and [-SLL,VHL-,-KLA] (from the lower alignment)

In this example, alignments (i) and (ii) are the same, that is [AHL-,SLL-,VHL-,-KLA]. Because this alignment is evaluated as the best, we select it as the alignment for the inconsistent region and we get the result shown in Step 2. The final alignment is produced by exchanging '.' wirh '-' (Step 3).

Now that we have finished core alignment, all we have to do is to attach remaining sequences one by one in the order decided by "Cluster Analysis". In this version of MASCOT, DP is used to attach each sequence. When we finish attaching all remaining sequences, the alignment we have is initial alignment.

It is not often the case, but sometime the attachment of a sequence to the result of 3-way DP may have better evaluation than the alignment by 3-way DP merging method. This possibility is also checked by MASCOT and the higher evaluated one is selected as the core alignment.

4 Refinement of alignment by Simulated Annealing

Initial alignment generated by MASCOT is of higher quality than that generated by conventional alignment systems. However, MASCOT improves the quality further by employing Simulated Annealing (SA) (Kirkpatrick *et al.*,1983) . We believe high quality alignment is necessary to extract motif information from alignment.

SA is a stochastic algorithm to search for the optimal answer, when evaluation of an answer is defined by some measure. SA introduces the notion of temperature to prevent the searcher of answer from being trapped near a local minimum in the search space. At low temperatures, the search is basically only permitted in the direction in which the evaluation value increases. While at high temperatures the search is also permitted in the direction that evaluation value decreases. SA can search for the globally optimal answer by gradually reducing temperature. This means that a temperature cooling schedule is necessary and important. At higher temperatures, the searcher can move to the vicinity of the globally optimal answer and then in the lower temperature. The globally optimal answer is sought in the small search space.

We formulated the problem of multiple alignment within the scheme of SA (Ishikawa *et al.*, 1991b). With SA, it is theoretically possible to start searching the optimal alignment from any initial alignment. However, this takes excessive amount of computational time even when the size of the problem is normal. Therefore, it is effective to apply SA to half-aligned alignments (Ishikawa *et al.*, 1991b).

To start SA from a half-aligned state, it is necessary to reduce the temperature gradually from moderate temperature. This means that it is imporant to have a cooling schedule suitable to the size and nature of sequences to be aligned. MASCOT copes with this problem by using *temperature parallel SA* (Kimura and Taki,1990) that assigns different temperatures to individual processors of parallel machine. In the scheme of temperature parallel SA, each processor searches for the optimal answer respectively at the temperature assigned to it. Each processor exchanges its answer with the processor with the adjacent temperature in a probabilistic manner. The exchange is executed at some regular intervals. Consequently, the best answer at the moment (semi-optimal answer) is found in the processor of the lowest temperature.

In this way, temperature parallel SA generates semi-optimal answers without requiring a cooling schedule. If we continue this SA infinitely, the optimal answer could be obtained with an appropriate cooling schedule. However when a half-aligned alignment is supplied to SA, the optimal answer is produced quickly with high probability.

5 Cluster Analysis

MASCOT produces multiple alignment based on 3 way DP. However, to produce high quality multiple alignment, a highly reliable alignment of three sequences must be selected A method for evaluating reliability of alignment of three sequences is necessary.

In contrast to the cluster analysis invented so far to make alignment, MASCOT makes cluster analysis by using an index of the similarity of three sequences in stead of that between two sequences. MASCOT uses the result of analysis based on this index throughout the process of alignment: selection of the core of cluster, deciding of the order of attachment, and deciding the order in which to combine clusters.

We introduce index R(i,j,k) which corresponds to the similarity of sequence i, sequence j, and sequence k. This index is calculated as follows. Firstly S(p,q), the similarity index between sequence p and sequence q, is calculated. In this version of MASCOT, S(p,q) is the evaluation of the alignment of sequence p and sequence q obtained by DP. Then R(i,j,k) is computed by summation of S(i,j), S(j,k) and S(k,i) (We are investigating the feasibility of weighted summation of S(p,q).). Because an alignment of similar sequences generally produces a reliable alignment, R(i,j,k) can be regarded as the reliability of the alignment of these three sequences obteined by DP.

After R(i,j,k) of all of the triplet {i,j,k} is calculated, these triplet is sorted into a list in decreasing order of R(i,j,k). Then, all the preparation for clustering is completed.

The clustering is iteration in three steps. The first step is to fetch one triplet from the top of the list. The second step is to check whether the three sequences in the triplet are included in some cluster or not. The third is to do the following:

(1) If none of three sequences are registered in any cluster : Make a new cluster, and register this triplet as the core of the cluster. We call this *core registration.*

(2) If two sequences of the triplet are registered in a cluster and the remaining sequence is not registered in any cluster : Register the unregistered sequence in the cluster where the other two are registered. We call this *attachment registration.* The special case is that the cluster has three sequences only. In that case, MASCOT doesn't attachment-register the sequence, but core-register the sequence in addition to the three sequences that are already core-registered.

(3) If two sequences of the triplet are registered in a cluster and the remaining sequence is registered in another cluster : Register the combining of these two clusters. We call this *combining registration.*

(4) Otherwise : Do nothing and discard the triplet.

Naturally the first triplet of the list falls into **(1)** and is core-registered because there is no cluster at the beginning. Succeeding triplets fall into **(1)** or **(2)** with high probability. And some clusters are generated by core registration and are developed while some are generated by attachment registration. The history of core-registration and attachment registration is transmitted to "Intra-cluster Alignment". There core registration information is used in core alignment and attachment registration information is used in attachment of alignment.

When a few sequences are left unregistered, triplets corresponding to **(3)** begin to appear. Combining registration information produced by **(3)** is transmitted to "Inter-cluster Alignment" where clusters are combined as prescribed in the information. If sequences to be attached appeared after some execution of "Inter-cluster Alignment", these sequences will be attached to clusters that are already inter cluster aligned.

6 An Example of Application of MASCOT

Sixteen protein sequences corresponding to endonuclease of retro-virus are aligned by MASCOT. These sequences are obtained from database PIR. The results of "Cluster Analysis" are shown in Fig.3 and the alignment according to this cluster analysis is shown in Fig.4. Each tiny circle in Fig.3 represents some sequence. Rectangles in Fig.3 indicate core alignment, oblongs inside a cluster indicate attachment-alignment, and oblongs outside cluster indicate intra-cluster alignment.

Firstly, sequences were divided into three clusters by "Cluster Analysis". Secondly, each sequences belonging the same cluster are aligned by "Intra-cluster Alignment". Though an initial alignment of Cluster 1 didn't change by 1 hour refinement with SA, the initial alignments of Cluster 2 and Cluster 3 were refined.

Thirdly, alignments of Cluster 2 and Cluster 3 were combined by "Inter-cluster Alignment". Finally, the resultant alignment and the alignment of Cluster 1 were combined. The sequences of Cluster 1 are sequences of lentivirinae and almost all sequences of Cluster 2 and Cluster 3 are sequences of oncovirinae.This indicates that clustering was effective.

This alignment result shows that the core alignment of each cluster has some clear sequence pattern and the alignment of all sequences of the cluster has vague but discernible patterns. We think this observation indicates the effectiveness of the strategy where the first step is to make reliable core alignment and the second step is to attach the remaining sequences.

All that 16 aligned sequences have a motif of ''H H C C'' . It can be regarded as some class of Zinc Finger motif. It is well-known that proteins with Zinc Finger can have the function to bind DNA. The ability of MASCOT to detect such a biologically important portion of protein shows the possibility that MASCOT can be developed to predict the functions and structure of unknown proteins.

7 Discussion

As indicated before, the result of alignment(Fig.4) shows clear consensus patterns in core alignments and vague but discernible patterns in the alignment of each cluster. We think that it is promising way to compare this kinds of pattern information with known motif information so that integrated information can be useful when attachment-alignment and intra-cluster alignment are performed.

We are now investigating how to use knowledge engineering to realize such an extension of MASCOT.

8 Summary

Firstly, an overview of multiple alignment system of protein, MASCOT, was introduced. Secondly, three characteristics of MASCOT, namely, (1) Generation of Initial Alignment by using 3-way DP to make the core of alignment, (2) Refinement of alignment by Simulated Annealing, (3) Cluster analysis of sequence by using triplets of sequences as units. Finally, an example of application of MASCOT to retro-viral sequences was shown and the identification of biologically important portion of sequences was indicated.

Acknowlodgements

The authors would like to especially thank Dr.Kanehisa and Dr.Nitta for valuable discussions. We would also like to thank members of the Genetic Information Processing Working Gruop at ICOT.

References

Barton,J.G. (1990) Protein Multiple Alignment and Flexible Pattern Matching. in *Methods in Enzymology Vol.183*, Academic Press, 626-645.

Dayhoff,M.O., Schwatz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change inproteins. In Dayhoff,M.O.(ed), *Atlas of Protein Sequence and Structure Vol.5, Suppl.3*, Nat. Biomed. Res. Found., Washington, D. C., 363 373.

Hirosawa,M., Hoshida,M., Ishikawa,M. and T. Toya,T. (1991) Multiple Alignment System for Protein Sequences employing 3-dimensional Dynamic Programming. *Genome Informatics Workshop II*, (in Japanese).

Ishikawa,M., Hoshida,M., Hirosawa,M., Toya,T., Onizuka,K. and Nitta,K. (1991a) Protein Sequence Analysis by Parallel Inference Machine. *Information Processing Society of Japan, TR-FI-23-2*, (in Japanese).

Ishikawa,M., Toya,T., Hoshida,M., Nitta,K., Ogiwara,A. and Kanehisa,M. (1991b) Multiple Alignment by Parallel Simulated Annealing. *Genome Informatics Workshop II*, (in Japanese).

Kimura,K. and Taki,K. (1990) Time-homogeneous Parallel Annealing Algorithm. in *Proc. Comp. Appl. Math. 13 (IMACS'91)*, 827-828.

Kirkpatrick,S., Gelatt,S.D. and Vecci,M.P. (1983) Optimization by Simulated Anealing. *Science, Vol.220, no.4598.*

Nakajima,K., Inamura,Y., Ichiyoshi,N., Rokusawa,K. and Chikayama,T. (1989) Distributed Implementation of KL1 on the Multi-PSI/V2. *Proc. 6th Int. Conf. on Logic Programming.*

Needleman,S.B. and Wunsch,C.D. (1970) A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins. *J. of Mol. Biol.*, **48**, 443-453.

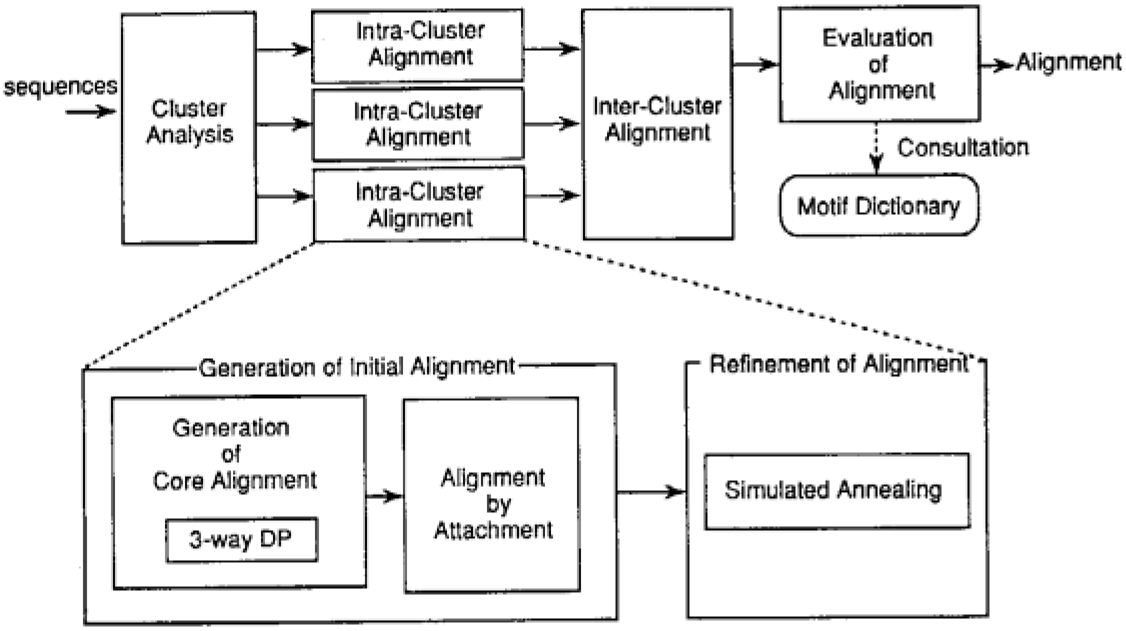

Fig.1. Multiple Alignment System MASCOT : Firstly, sequences are grouped into clusters by Cluster Analysis. Secondly, the sequences belongs to same cluster are aligned by Intra-Cluster Alignment. Thirdly, clusters are combined by Intra-Cluster Alignemnt. Finally, resultant alignments are evaluated by consulting motif dictionary to select optimal alignment.

In Inter-Cluster Alignment, two-phase alignment is employed. In the first phase, an inital alignment is generated based on 3-way DP. In the second phase, the alignment is refined by Simulated Annealing.

```
Step0   HTLV-1   : LLQAIAHLGKPSYINT        Step1     HTLV-1   : LLQAIAHL.GK.PSY.INT
        HTLV-2   : VLQAISLLGKPLHINT                  HTLV-2   : VLQAISLL.GK.PLH.INT
        BLV      : LLEAIVHLGRPKKLNT                  BLV      : LLEAIVHL.GR.PKK.LNT
                    ^ ^^  ^^ ^   ^^                             1111100001111111111
        HTLV-2   : VLQAI-SLLGK-PLH-INT               HTLV-2   : VLQAI-SLLGK-PLH-INT
        BLV      : LLEAIVHL-GR-PKK-LNT               BLV      : LLEAIVHL-GR-PKK-LNT
        HIV      : FL--L-KLAGRWPVKTIHT               HIV      : FL--L-KLAGRWPVKTIHT
                    ^     ^ ^  ^     ^

Step2   HTLV-1   : LLQAIAHL.GK.PSY.INT     Step3     HTLV-1   : LLQAIAHL-GK-PSY-INT
        HTLV-2   : VLQAISLL.GK.PLH.INT     (Result)  HTLV-2   : VLQAISLL-GK-PLH-INT
        BLV      : LLEAIVHL.GR.PKK.LNT               BLV      : LLEAIVHL-GR-PKK-LNT
        HIV      : FL--L-KLAGRWPVKTIHT               HIV      : FL--L-KLAGRWPVKTIHT
                    ^     ^ ^  ^     ^                          ^     ^ ^  ^     ^
```

Fig.2. Generation of Core Alignment by 3-way DP Merging Method : By using 3-way merging method, core aligument of 4 sequences is generated by combing two 3-way alignment produced by 3-way DP. Two alignments have two common sequences, namely, HTLV-2 and BLV(Step 0). In Step 1, by comparing two aligunents of those 2 sequence, the alignments are dedided into consistent regions (marked by '1') and inconsistent regions (marked by '0'). In Step 2 and Step 3, core alignment is constructed by using the imformation of the consistent regions and the inconsistent regions.

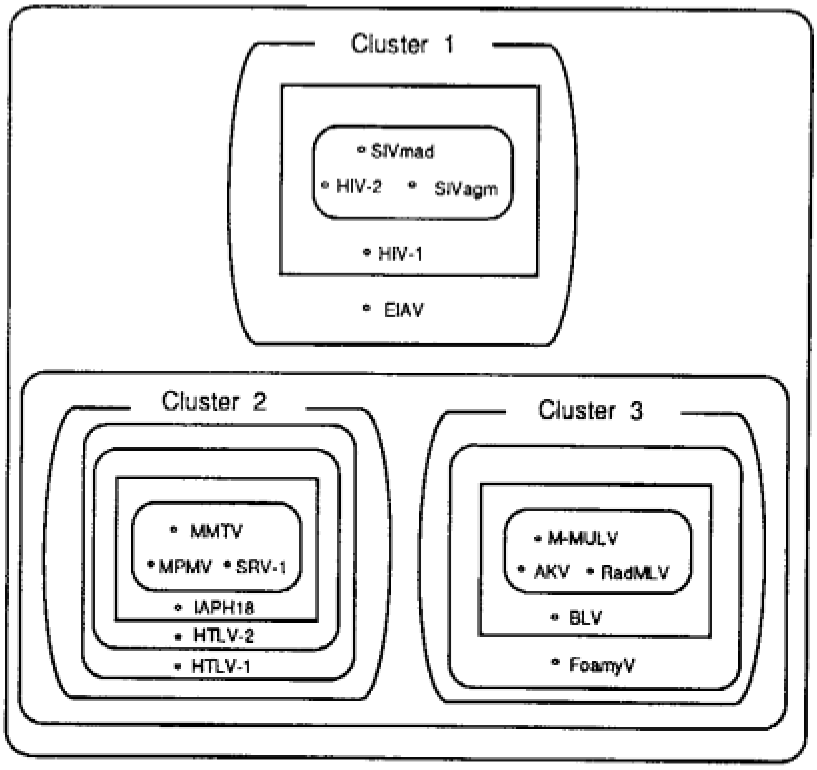

Fig.3. Results of Cluster Analysis : Sequences are grouped into three clusters. Cluster 2 and Cluster 3 are combined to form bigger cluster.

```
<Cluster1>
HIV-2 : AHKGIGG--NQEVDH-L-VSQGI-RQVLF-L-----EK-IE-PAQEEHEKYHS---NVKELSHK-----
SIVmad: AHKGIGG--NQEIDH-L-VSQGI-RQVLF-L-----EK-IE-PAQEEHSKYHS---NIKELVFK-----
SIVagm: AHKGIGG--NEEIDK-L-VSKGI-RRVLF-L-----EK-IE-EAQEKHERYHN---NWKNLADT-----
HIV-1 : AHKGIGG--NEQVDK-L-VSAGI-RKILF-L-----DG-ID-KAQDEHEKYHS---NWRAMASD-----
EIAV  : ENK-I----SQRGDKGF-GSTGV-----FWV-----EN-IQ-EAQDEHENWHT---SPKILARN-----
        ahKgIGG  N   D  l VS Gi r  lF L        I   AQ  H  yH    n

<Cluster2>
SRV-1 : AHSGLPG-PIAHGNQ-K-A-D-LATK--T-VASNINTN-LE-SAQNAHTLHHL---NAQTLKLM-----
MPMV  : AHSGLPG-PIAQGNQ-R-A-D-LATK--I-VASNINTN-LE-SAQNAHTLHHL---NAQTLRLM-----
MMTV  : GHTGLPG-PLAQGNA-Y-A-DSL-TR--I-L-----TA-LE-SAQESHALHHQ---NAAALRFQ-----
IAPH18: AHSGLPG-PMSLGND-L-A-D-KATK--L-VATALSTH-AQ-AAKEFHKRFHV---TAETLRRR-----
HTLV-2: SHTNLPD-PISTFNE-Y-T-DSL-----I-LAP-L-VP-L--TPQGLHGLTHC---NQRAL-VS-----
HTLV-1: SHTNLPD-PISRLNA-L-T-DAL-----L-I-----TPVLQLSPAELHSFTHC---GQTALTLQ-----
         H gLPg P   gN    a D               t      a   H   H    a  L

<Cluster3>
AKV   : GYWVFQGKPVMP-DQ-F-VFE-------L-L-----DS-L-------HRLTHLGYQKMKAL-LDRGESP
RadMLV: GYWVFQGKPVMP-DQ-F-VFE-------L-L-----DS-L-------HRLTHLGYQKMKAL-LDRGESP
M-MULV: KYWVYQGKPVMP-DQ-F-TFE-------L-L-----DF-L-------HQLTHLSFSKMKAL-LERSHSP
BLV   : SHSS-ASHPIASLNN-Y-VDQ-------L-L------P-LE-TPEQWHKLTH--CNS-RAL-SRWPNPR
FoamyV: VSRP-EGVKIIP-PQ-S-DRQKI-----V-L-----QA---------HNLAHTG--REATL-LKIAN--
                p                   l L        l       H LtH       aL


<Cluster1>
HIV-2 : -F---GIPNLVARQ---IVNSCAQCQQ---------KG-EA-I-HGQVNAELGT-WQ---MDCTH-LEG
SIVmad: -F---GLPRLVAKQ---IVDTCDKCHQ---------KG-EA-I-HGQVNSDLGT-WQ---MDCTH-LEG
SIVagm: -Y---GLPQIVAKE---IVAMCPKCQI---------KG-EP-V-HGQVDASPGT-WQ---MDCTH-LEK
HIV-1 : -F---NLPPVVAKE---IVASCDKCQL---------KG-EA-M-HGQVDCSPGI-WQ---LDCTH-LEG
EIAV  : -Y---KIPLTVAKQ---ITQECPHCTK---------QG-SG-P-AGCVMRSPNH-WQ---ADCTH-LDN
              lP  VA     Iv  C  C           kG e    hGqV    g  WQ    DCTH Le

<Cluster2>
SRV-1 : -F---NIPREQARQ---IVRQCPICATYLPVPH-L--G-VN-P-RGLL--PNMI-WQ---MDVTH-YSE
MPMV  : -F---NIPREQARQ---IVKQCPICVTYLPVPH-L--G-VN-P-RGLF--PNMI-WQ---MDVTH-YSE
MMTV  : -F---HITREQARE---IVKLCPNCPDWGHAPQ-L--G-VN-P-RGLK--PRVL-WQ---MDVTH-VSE
IAPH18: -F---ALSRKEARE---IVTQCQNCCEFLPTPH-M--G-IN-P-RGIR--PLQM-WQ---MDVTH-IPS
HTLV-2: -F---GATPREAKS---LVQTCHTCQTINSQHH-MPRG-YI-R-RGLL--PNHI-WQ---GDVTH-Y-K
HTLV-1: -----GATTTEASN---ILRSCHACRGGNPQHQ-MPRG-HI-R-RGLL--PNHI-WQ---GDITH-FK-
         F      r  Ar    iV  C  C      p     G  n p RG    P    WG   mDvTH

<Cluster3>
AKV   : YY---MLNRD--KTLQYVADSCTVCAQVNASKAKIGAG-VR-V-RGHR--PGSH-WE---IDFTE-VKP
RadMLV: YY---MLNRD--KTLQYVADSCTVCAQVNASKAKIGAG-VR-V-RGHR--PGTH-WE---IDFTE-VKP
M-MULV: YY---MLNRD--RTLKNITETCKACAQVNASKSAVKQG-TR-V-RGHR--PGTH-WE---IDFTE-IKP
BLV   : -----ISAWD-PRS---PATLCETCQKLNPT----GGGKMRTIQRG-W--APNHIWQ---ADITH-YK-
FoamyV: LYWWPNMRKDVVKQL----GRCQQCLITNASNK--ASG-PI-L-RPDR--PQKP-FDKFFIDYIGPLPP
                 D           C  C   N        G      Rg       h w     D t   k
```

Fig.4. Example of Alignment using MASCOT : This figure shows the result of alignment of some portion of retro-viral sequences of endonuclease. All the sequences were obtained from *PIR*. The upper four sequences of each cluster constitute core sequences. Consensus residues are indicated below the alignment of each cluster . The uppercase letters designate consensus throughout all sequences of the cluster and the lowercase letters designate consensus among core sequences only.