DNA^{+Pro}: Combined DNA-Protein Sequences Improve Alignment and evolutionary Analysis of Protein and their-coding DNA sequences

Xiaolong Wang[1*]

1. Department of Biotechnology, Ocean University of China, Qingdao, 266003,

People's Republic of China,

* E-mail: xiaolong@ouc.edu.cn

ABSTRACT

Alignment of DNA and protein sequences is the basis of evolutionary, structural and functional studies. Present sequence alignment methods are error-prone, producing systematical bias. Errors in sequence alignments lead to subsequent misinterpretation of evolutionary, structural and functional information in genes, proteins and genomes. In traditional sequence alignment algorithms, aligning of DNA sequences and protein sequences are conducted separately. And it has been long believed that the phylogenetic signal disappears more rapidly from DNA sequences than from the encoded proteins. It is therefore generally preferable to align coding DNA sequence at the amino acid level. Here we present a new method that combines DNA and protein sequences in a single alignment, and gives more accurate alignments for both DNA and protein sequences. By analyzing primate lentivirus proteins and *Staphylococcus aureus* restriction enzyme and their coding DNA sequences, we demonstrated that combined DNA-protein sequences improve the quality of multiple sequence alignment and the analyses of practical evolutionary problems. We further showed that the theoretical information content increases when DNA and protein sequences is combined, and calculation of distances of aligned combined DNA-protein sequences are more biological meaningful, thus improves the accuracy of progressive sequence alignment and phylogeny analysis by integrating sequence information, which is buried separately in DNA and protein sequences.

INTRODUCTION

Bioinformatics, as well as functional and comparative genomics, seek to discover functional and structural sequence changes leading to genetic differences between species, as well as to provide accurate reconstruction of evolutionary histories of related genes, proteins and genomes. However, all evolutionary, structural or functional studies that rely on sequence analyses require accurate sequence alignments, *i.e.*, the correct identification of homologous nucleotides or amino acids, and the accurate positioning of gaps indicating insertions and deletions. For example, progressive algorithms, such as ClustalW (1), build a multiple alignment from pairwise alignments between sequences, performed in order of decreasing relatedness according to a guide tree. Unfortunately, although there are quite a few sequence alignment algorithms available, sequence alignment is still highly error-prone. Different sequence alignment tools often lead to drastically different conclusions in phylogenetic analyses, and can support entirely different mechanisms driving evolutionary, structural and functional changes in sequences. Here we present a new method, DNA^{+Pro}, which combines DNA and protein sequences in a single alignment, thus prevents some of these errors, improves the quality of sequence alignment and the accuracy of phylogeny analysis.

RESULTS

Combined DNA-protein sequence improves multiple sequence alignment

As shown in Fig 1A, the traditional protein-only alignment aligned by ClustalW, and its combined- or DNA-view reverse-translated by DNA^{+Pro}, suggests that part of the variable (V_2) region has a high rate of amino acid and DNA base substitutions. Obviously, as shown in Fig S1A, ClustalW incorrectly squeezed ‘homology’ insertions and non-insertions between highly conserved sequence blocks. Using phylogeny-guided heuristic iteration, alignments aligned by MAFFT (Fig S1B), MUSSLE (Fig S1C) and T-coffee (Fig S1D) are improved to some degree, but the problem still exists, and insertions and non-insertions are still mismatched to each other.

With a phylogeny-aware algorithm (*2*) that considers the phylogeny and distinguishes insertions from deletions, PRANK provides a very different alignment of this region (Fig S1E). The PRANK algorithm keeps insertions “flagged” to ensure that independent insertions are not matched. The PRANK algorithm correctly identified insertions, while produces too many gaps, and ignores true homology among insertions. The PRANK algorithm, therefore, provides only a partial solution to the erroneous treatment of insertions in progressive alignment.

These errors are avoided by aligning combined DNA-protein sequences using DNA^{+Pro} algorithm. The principle and implementation of the DNA^{+Pro} algorithm is described in the material and method section. Protein sequences and their coding DNA sequences were convert into combined DNA-protein sequences, in which every triplet codon is immediately followed by the one-letter code of its encoded amino acid. The combined sequences are then aligned using progressive algorithm by calling CLUSTAL W with a *combined DNA-Protein (CDP) scoring matrix*, such as CDP-Gon250 matrix (Table S1),

and a set of user-defined settings (Table S2). A combined alignment of HIV and SIV gp120 DNA and protein sequences are shown in Fig 1B, the combined alignment can be visualized in combined-view, protein-view or DNA-view. As shown in Fig 1B, DNA^{+Pro} not only identify the distinct insertions and show the homology among insertions at the same time, but also give more accurate alignment in the conserved region (Fig S1F).

Combined alignment improves phylogenic analyses of HIV and SIV proteins

Traditional sequence alignment algorithms perform heuristics pairwise alignments at the branching points of a guide phylogenetic tree approximating the evolutionary history of the sequences. An input guide tree is rarely consistent with the output phylogenetic tree constructed from a multiple sequence alignment, and the output tree is often inconsistent with the real evolutionary history. Consequently, the output phylogenetic tree and the alignment are both highly error-prone. As shown in Fig S2, the phylogenetic tree of the *env* gene inferred from the protein-only alignments by CLUSTAL W (Fig S2A), MAFFT (Fig S2B), T-coffee (Fig S2C), MUSSLE (Fig S2D) and PRANK (Fig S2E) are all varied markedly, and the tree inferred from the combined alignment suggests a different evolutionary process (Fig S2F).

Moreover, it has been suggested that different regions in a HIV or SIV genome have different evolutionary histories. By systematically examining trees from all possible combinations of four SIVs in four genomic regions, it was inferred that the chimpanzee SIV virus (SIVcpz) is mosaic: Its left-hand region (*gag* and *pol*) comes from a red-capped mangabey virus, and the right-hand region (*env*) is the ancestor of a virus found in several Cercopithecus monkeys (*3*). The mosaic structure of SIVcpz requires that a chimpanzee was infected with two different monkey viruses and these recombined. It was speculated that the recombination may occur in a dually infected monkey and the mosaic virus was transmitted to a chimpanzee, but it is more probable that the dual infection occurred in a chimpanzee since chimpanzees hunt and eat small monkeys (*3*).

Since HIV is origin from SIVcpz, it is interesting to know whether HIV genomes are also mosaic, and whether such kind of dual infection and recombination had also happened in human. Bootscan analysis, which breaks the genome into small sections and analyzes each section independently, has been used to identify areas of recombination within an HIV-1 genome (*4, 5*). However, it is unclear whether inconsistent bootstrap support for phylogenetic incongruence is sufficient evidence of the presence of recombinant lineages (*5*). The apparent phylogenetic incongruence at different regions of the genome that was taken as evidence of recombination is shown to be not statistically significant (*5*). Furthermore, simulations indicate that bootscanning and pairwise distance results, previously used as evidence for recombination, can be misleading, particularly when there are differences in substitution or evolutionary rates across the genomes of different subtypes (*5*). A likely explanation for the differences in the evolutionary rates across the genome is that different regions of the genome are under different selective pressures (*5*).

We performed phylogeny analysis for HIV *env* and *gag* genes using protein-only and combined alignments respectively. As shown in Fig 3A, the phylogenetic tree for *env* and *gag* constructed from protein-only alignments are varied, look as if some of the HIV

genomes, such as HV1J3, HV1B1, HB1A2, HV2BE and HV2G1, are mosaic. However, the trees inferred from the combined alignments (Fig 3B) show a nearly completely consistent evolutionary process both for *env* and *gag,* suggests that different regions of most HIV genome underwent similar evolutionary histories since isolated from SIVcpz, and therefore dual infection and recombination had never or rarely happened in these HIV strains. We believed that these trees constructed from combined alignments are more accurate than the other ones, because not only they are consistent for different genes, but they have much higher Bootstrap values (Fig 2). In fact, they also have a stronger and more biological meaningful theoretical basis, as described in the material and method section. Artifacts of traditional protein-only or DNA-only alignment algorithms may cause misinterpretation of a 'mosaic structure' in closely-related genomes, and the inappropriate attribution of recombinant origins to divergent sequences obscures the true evolutionary properties of these viruses (*5*). DNA^{+Pro} provide a more accurate analysis tool that can prevent this kind of errors.

Combined alignment improves interpretation of evolutionary events

With the aid of the phylogenetic tree inferred from the combined alignment, and the 64-color views of the combined alignment (Fig 1A), one can easily interpret the mutation events that happened in the evolution process of this region. For example, three insertions, GGNSSNGNGDSSK, EKGNISPKNNTSNNTS and NNSTKDNIKNDNST, were identified respectively in HV1ZH, HV1RH and HV1J3. According to the phylogenetic tree, HV1ZH is the ancestor of HV1RH and HV1J3, from the combined-view of the alignment, we can speculate that the mutation events for HV1ZH to evolve were: in HV1RH, except for some base substitutions and a one-residue insertion (K) in the left, the right part is replaced by a repetitive sequence (PKNNTSNNTS); and then in HV1J3, except for some base substitutions and a two-residue deletion (SP) in the middle, one of the two tandem repeats (NNST) shifted from the right to the left.

Moreover, as shown in the combined- and the DNA-view of the combined alignment, the coding sequences of these insertions are changing from GC-rich to more and more AT-rich and repetitive. Comparing the combined-view with the protein-view and the DNA-view, it becomes clear that insertions and deletions (Indels) in this variable region might be caused by slipped strand mispairing: after the virus has inserted the two copies of its RNA genome packaged in the virion into the host's cell, the viral reverse transcriptase, encoded by the *pol* gene, reverse transcribes the RNA to DNA. As the polymerase progresses it hops from one copy of the genome to the other (*3*). Indels will easily occur in the variable regions of the HIV genomes. Considering the recombination feature of HIV reverse transcriptase, and the slipped strand mispairing property of short tandem repeats, this mechanism is more convincing when compared with mutation events that were suggested by other alignments, such as a lot of amino acid substitutions (Fig 2A-2D), or distinct insertions at the same position (Fig 2E)(*2*). As shown in the combined alignments of the *env*, *gag* and *pol* genes (Supplementary supporting material), such kind of indels happened again and again in the other HIV strains, and in other variable regions. Therefore, slipped strand mispairing might be the major cause of highly frequent Indels in the variable region of HIV genomes.

Combined alignment improves phylogenic analyses of restriction enzymes

We aligned and constructed phylogenetic tree for a new class of restriction enzymes (SAUSA300_2431) and their coding DNA sequences from Gram positive bacteria *Staphylococcus aureus*, respectively with ClustalW and DNA^{+Pro}. Comparing to the phylogenetic tree constructed from ClustalW Protein-only alignment (**Fig 3A**), the tree from DNA^{+Pro} combined alignment (**Fig 3B**) is more consistent with the robust multi-gene phylogenetic tree for *Staphylococcus aureus subsp. aureus* (**Fig 3C**). For example, enzyme from strain MRSA252 is grouped as a common ancestor in the combined tree, but misplaced in a subgroup in the protein-only tree.

Information content of DNA, protein and combined sequences

The theoretical measurements of information content are given by Shannon entropy. For DNA and protein sequences, the information contents are determined respectively by the frequencies of bases or amino acids occurred in a sequence:

$$S_{DNA}=-\Sigma_{i=1,4}\, P_i \ln(P_i)$$

$$S_{Pro}= -\Sigma_{j=1,20}\, P_j \ln(P_j)$$

The information contents of DNA and protein sequences reach their lower limit when the sequences are random (bases and amino acids occur in equal frequencies):

$$\text{Min}\,(S_{DNA}) = -4 \times (1/4 \times \ln\,(1/4)) = -\ln\,(1/4) = 1.386$$

$$\text{Min}\,(S_{Pro}) = -20 \times (1/20 \times \ln\,(1/20)) = -\ln\,(1/20) = 2.996$$

The minimal information content of protein sequences is larger than that of DNA sequences, forms the basis that ensures protein sequences are more informative than DNA, and outperforms in sequence alignment.

In a combined DNA-protein sequence, an amino acid is uniquely determined by the three bases of its encoding triplet codon, so the entropy is dependent on the frequencies of the triplet codons, but independent on the frequencies of the amino acids. Let P_k, P_l, P_m be frequencies of three bases of a triplet codon. The theoretical information content of a combined DNA-protein sequence is given by

$$S_{DNA+Pro}= -\Sigma_{\,k,\,l,\,m=1,\,4}\, P_k P_l P_m \ln\,(P_k P_l P_m)$$

The minimal information content of combined sequences is reached when and only when the DNA bases occur in equal frequencies:

$$\text{Min}\,(S_{DNA+Pro}) = -64 \times (1/64 \times \ln\,(1/64)) = -\ln\,(1/64) = 4.159$$

The minimal information content of combined DNA-protein sequences is larger than those of DNA and protein sequences. DNA^{+Pro} compute information contents for DNA, protein and combined sequences. It is shown that computed information contents of combined sequences are always greater than those of corresponding protein sequences or DNA sequences, and this forms the theoretical basis of the combined alignments.

DISCUSSION

Problems in DNA-only or protein-only alignments

The small size of the alphabet makes alignment of nucleotide sequences inherently difficult: even a pair of completely unrelated DNA sequences will typically display ~25% identity over their entire length and it is often possible to find extended local alignments

where >50% of the aligned nucleotides are identical. This makes the task of distinguishing true homology from random similarity difficult. The simple fact that proteins are built from 20 amino acids while DNA only contains four bases, means that the 'signal-to-noise ratio' in protein sequence alignments is better than in DNA sequence alignments (*6*). Besides this theoretical information-content advantage, protein sequence alignments also benefit from amino acid substitution matrices, such as BLOSUM, PAM and Gonnet series. These matrices contain empirically derived scores for each possible amino acid pair and provide a rational basis for aligning amino acids (*6*).

Taken together with generally higher rates of synonymous mutations over non-synonymous ones, it has long been believed that the phylogenetic signal disappears much more rapidly from DNA sequences than from the encoded proteins. It is therefore generally preferable to align protein coding DNA sequences at the amino acid level. For example, some programs, such as RevTrans (*6*), construct a multiple DNA alignment by translating the protein coding DNA sequences, aligning the resulting peptide sequences, and building a multiple DNA sequence alignment by reverse translating of the aligned protein sequences. But some important information carried by DNA sequences, such as synonymous mutations and frame-shift events, get lost after they were translated into protein sequences. An interesting coding sequence alignment algorithm, COMBAT (*7, 8*), combines DNA and protein sequence alignment. However, it seems that the algorithm is over complicated, and so far has not been used for multiple sequence alignment, but only for pairwise sequence alignment. Here by using combined DNA-protein scoring matrices, allows aligning multiple combined DNA-protein sequences in a single alignment, and overcomes the problems commonly exist in DNA-only or protein-only alignments.

Problems in multiple sequence alignment algorithms

Conventional progressive algorithms (*9–18*) always match neighboring insertions and non-insertions in the same column if they produced significant homology. This problem has been identified as being caused by repeated penalizing gap-opening (*19*), but cannot be avoiding by reducing the gap-opening penalties in traditional protein-only alignment. If small gap-opening penalties are used, protein-only alignment algorithms will result in 'gappy' alignments. Recently, Ari Löytynoja, *et al* (*2*) uses the phylogeny-aware (PRANK) approach that "flags" the gaps made in previous alignments and, so that distinct insertion events are kept separate even when they occur at exactly the same position. The PRANK method, however, produces many unnecessary gaps, and ignores homology between insertions (Fig 3E). The authors argued that inserted characters are not descendants of any other insertions or ancestral characters, and thus should not be aligned with anything (*2*). However, if an inserted sequence is homologous to another insertion, it is possible that this insertion is the ancestor of the other one, so their homology still should be shown in the alignment, and proper handling of these insertions will affect the correctness of downstream phylogeny analysis, as well as structural and functional studies.

In conventional protein-only alignments, it is almost impossible to distinguish an insertion from a non-insertion if they are 'homologous' to each other in the amino acid level, because there is no information available to discriminate them. Using combined DNA-protein sequences, this problem becomes much easier to solve because of the

enlarged alphabet and the subsequently increased information content. In a combined DNA-protein sequence, every information unit consist three bases and one amino acid. In a combined alignment, triplet codons followed by its encoded amino acids shows detailed mutation events, such as the synonymous and non-synonymous base substitutions. And in DNA^{+Pro}, these mutations are clearly shown by 64-color views. In a highly conserved region, a rate of synonymous mutations is dominantly higher than that of non-synonymous ones. In a variable region, however, the rate of non-synonymous mutations is higher than in a conserved region. These correlated DNA bases and amino acids, together with the 64-color views, make it much easier to distinguish a base substitution from an insertion or deletion, and therefore easier to distinguish an insertion from another distinct insertion or non-insertion. Although sequences are 'homologous' to each other in a protein-only alignment (Fig 1A, top), they show significant differences in the reverse-translated combined-view (Fig 1A, middle) or DNA-view (Fig 1A, bottom). Combining information buried respectively in DNA and protein sequences, DNA^{+Pro} enables distinguishing base substitution from insertion more easily, and allows aligning homologous sites more accurately, overcomes the problems of conventional DNA-only or protein-only alignment while avoid producing unnecessary gaps.

Problems in molecular phylogeny analysis

A traditional molecular phylogenetic tree was constructed either from a DNA-only or a protein-only alignment. Distances of sequences, however, are both underestimated. When a protein-only alignment is used to construct a phylogenetic tree, synonymous mutations are ignored, because they have no contribution to the amino acid substitutions, so the mutational events happened in DNA sequences are underestimated. When a DNA alignment reverse translated from the protein alignment was used instead, although both synonymous and non-synonymous mutations are counted, amino acid substitution events, however, are ignored. Though every amino-acid substitution has at least one non-synonymous base substitution occurred in the DNA sequences, non-synonymous mutations are treated without discriminating from the synonymous ones. So the biological effects of non-synonymous mutations are underestimated. Using combined alignments, not only all base substitutions are counted, but also synonymous and non-synonymous mutations are discriminated by change (or not) in the amino acid sequences. In addition to more accurate sequence alignment, the more biologically meaningful estimation of distances for aligned sequences forms a stronger basis for phylogeny analysis.

CONCLUSION

Multiple sequence alignment is of crucial importance in subsequent genomic analyses, such as phylogeny inference, structure modeling and detection of positive selection (20). Our analysis shows that errors in traditional protein-only or DNA-only alignment may lead to serious errors in evolutionary and comparative study of protein-coding genes. It may not the progressive algorithm itself is defective. Rather, correct alignment requires that information buried separately in DNA and protein sequences to be fully exploited in a combined manner. DNA^{+Pro} is useful in cases where a multiple alignment of coding sequence forms the basis for further investigations, such as phylogenetic, structural and

functional analysis of closely-related proteins, and this combined alignment method gives a more accurate picture of the mechanisms of protein evolution, and may also be useful in structure and functional analyses of their coding DNA sequence.

MATERIALS AND METHODS

DNA and protein sequences and online resources

Envelope glycoprotein gp120 (Env) and gag polyprotein (Gag) sequences for different strains of human immunodeficiency virus (HIV) were derived from the seed alignment of Pfam family pf00516. Their corresponding complete coding DNA sequences (CDS) were retrieved from GenBank. The coding sequences of gag for HV1W1, HV1BN, HV1ZH and HV2D2 are not available in the public nucleotide databases, so they are not included in the alignments. Restriction enzyme Sausa300_2431 (KEGG ID saa:SAUSA300_2431) and its homologs is derived from protein cluster CLSK903935, and their corresponding complete coding DNA sequences (CDS) were retrieved from Ensemble genome browser. A robust multi-gene phylogenetic tree for *Staphylococcus aureus subsp. aureus* was downloaded from the PathoSystems Resource Integration Center (PATRIC) (*21*). (http://patricbrc.vbi.vt.edu/portal/portal/patric/Phylogeny?cType=taxon&cId=282458).

Combined DNA-protein (CDP) scoring matrix

A CDP scoring matrix, such as CDP-Gon250 scoring matrix (Table S1), is a 24 x 24 array derived by merging a nucleotide substitution matrix with an amino acid substitution matrix, such as Gonnet250, Blosum62 and PAM250 scoring matrices. In a CDP matrix, substitutions between any pair of amino acids, or any pair of DNA bases, are allowed, but high penalties are given to 'substitutions' between a DNA base and an amino acid to prohibit mismatch of a DNA base to an amino acid during the alignment process. As shown in Table S1, The first line of a CDP matrix is the symbol set, lowercase letters (a, c, g, u/t) stands for DNA/RNA bases, uppercase letters (A/B, C/X, D, E, F, G/Z, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) stand for amino acids. Numbers are the 24 x 24 scoring matrix. Clustal W do not support case-sensitive alphabet, to avoid the confliction in codes for DNA bases and amino acids, DNA^{+PRO} changes the amino acid codes temporary in combined DNA-Protein sequences: replaces A with B, C with X, and G with Z. In resolving of the output combined alignment, DNA^{+PRO} changes the amino acid codes back to ordinary codes, write them in uppercase, and the DNA bases in lowercase letters.

Combined DNA-protein sequences Alignment

Using a homemade program, DNA^{+PRO}, the coding DNA sequences were translated using standard translation table SGC0 and converted into combined DNA-Protein sequences, in which every triplet codon is immediately followed by the one-letter code of the corresponding amino acid. DNA^{+PRO} is coded in Microsoft Visual Basic for Applications (VBA) and integrated with Microsoft EXCEL. The combined DNA-protein sequences were then subjected to progressive alignment by calling CLUSTAL W (v. 2.0.12) using a set of user-defined settings and a combined DNA-Protein (CDP) scoring matrices, CDP-Gon250 scoring matrix (Table S1). The optimal parameters for Clustal W to align the combined sequences were listed in Table S2. When finished, the output combined alignment is feed back to DNA^{+PRO} for final refinement, gap-only columns are removed, and gaps produced in a triplet codon is moved to the right of the codon to keep the combined alignment in 4-letter groups. Finally a combined DNA-protein alignment was visualized by DNA^{+PRO} in combined-, protein- or DNA-view (Fig 1A).

Protein-only sequence alignment

For protein-only sequence alignment, the multiple sequence alignment methods used are CLUSTAL W v. 2.0.12, MAFFT v. 5.861, MUSCLE v. 3.6, T-COFFEE v. 3.93 and PRANK. All programs were run with default settings. As shown in Fig 2A, DNA^{+PRO} reverse-translate a protein-only alignment into a combined DNA-protein alignment, so that protein-only alignments can be compared with combined alignments in a combined-view.

Phylogeny analysis

For phylogeny analysis, any sites containing alignment gaps or missing data are removed, and pairwise *p*-distance for each alignment was calculated. Phylogeny trees was constructed with neighbor joining method and bootstrapped with 1000 replicates. Bootstrap consensus trees were output for every alignment. Phylogeny trees were constructed from protein-only alignments, combined alignments, or combined alignments reverse-translated from protein-only alignments.

For a traditional protein-only alignment, distance between two aligned protein sequences (P_{AA}) is defined as proportion of amino acid sites that are different:

$$P_{AA} = D_{AA}/L_{AA} \quad (1)$$

Where D_{AA} is the number of amino acid sites that are different between the two aligned protein sequences; L_{AA} is number of valid common amino acid sites compared.

When the protein-only alignment is reverse-translated into a DNA alignment, distance between the two aligned DNA (P_{Base}) sequences is defined as proportion of DNA bases that are different:

$$L_{Base} = 3L_{AA} \quad (2)$$

$$P_{Base} = D_{Base}/L_{Base} = D_{Base}/3L_{AA} \quad (3)$$

Where D_{Base} is the number of DNA base sites that are different between the two aligned DNA sequences; L_{Base} is number of valid common DNA bases compared.

In a combined alignment aligned by DNA^{+Pro}, or reverse translated from a protein-only alignment, distance between the two combined DNA-protein sequences (P) is defined as proportion of DNA base and amino acid sites that are different:

$$L = (L_{AA} + L_{Base}) = 4L_{AA} \quad (4)$$

$$P = D/L = (D_{AA} + D_{Base}) / 4L_{AA} \quad (5)$$

Where L_{AA} and L_{Base} is respectively number of common DNA base sites and amino acid sites compared; D_{AA} and D_{Base} is respectively number of DNA base sites and amino acid sites that are different between the two aligned combined DNA-protein sequences.

Obviously, for a given number of base substitutions, say d_{Base}, in a given length of coding DNA sequence, $l_{Base} = 3l_{AA}$,

$$p_{Base} = d_{Base}/l_{Base} = d_{Base}/3l_{AA}$$

P reaches its minimum when all of the base substitutions are synonymous, *i.e.*, there is no amino acid substitution,

$$d_{AA,\,min} = 0 \quad (6)$$

$$p_{AA,\,min} = 0 \quad (7)$$

$$p_{min} = d_{Base} / 4l_{AA} = 3/4\ p_{Base} \quad (8)$$

On the other hand, P reaches its maximum when all of the base substitutions are

non-synonymous, and each amino acid substitution is caused only by one DNA base substitution, therefore,

$$d_{AA,\,max} = d_{Base} \quad (9)$$

$$p_{AA,\,max} = d_{Base} / l_{AA} = 3\,p_{Base} \quad (10)$$

$$p_{max} = 2d_{Base} / 4l_{AA} = 3/2\,p_{Base} \quad (11)$$

Usually, base substitutions may be either synonymous or non-synonymous, and each amino acid substitution is caused by one, two or three base substitutions, therefore,

$$0 \le P_{AA} \le 3P_{Base} \quad (12)$$

$$3/4\,P_{Base} \le P \le 3/2\,P_{Base} \quad (13)$$

According to equation (12), distances calculated from a protein-only alignment may seriously underestimate the distance of two coding DNA sequence when there is a high rate of synonymous mutations. On the other hand, the distance of two coding DNA sequence may seriously underestimate the distance of their encoded protein sequence when there is a high rate of non-synonymous mutations. Equation (13), however, suggests that distances calculated from combined alignments give better estimations of the distances for the coding DNA and their encoded protein sequences, as they dependent on not only rates of base substitution, but rates of amino acid substitution.

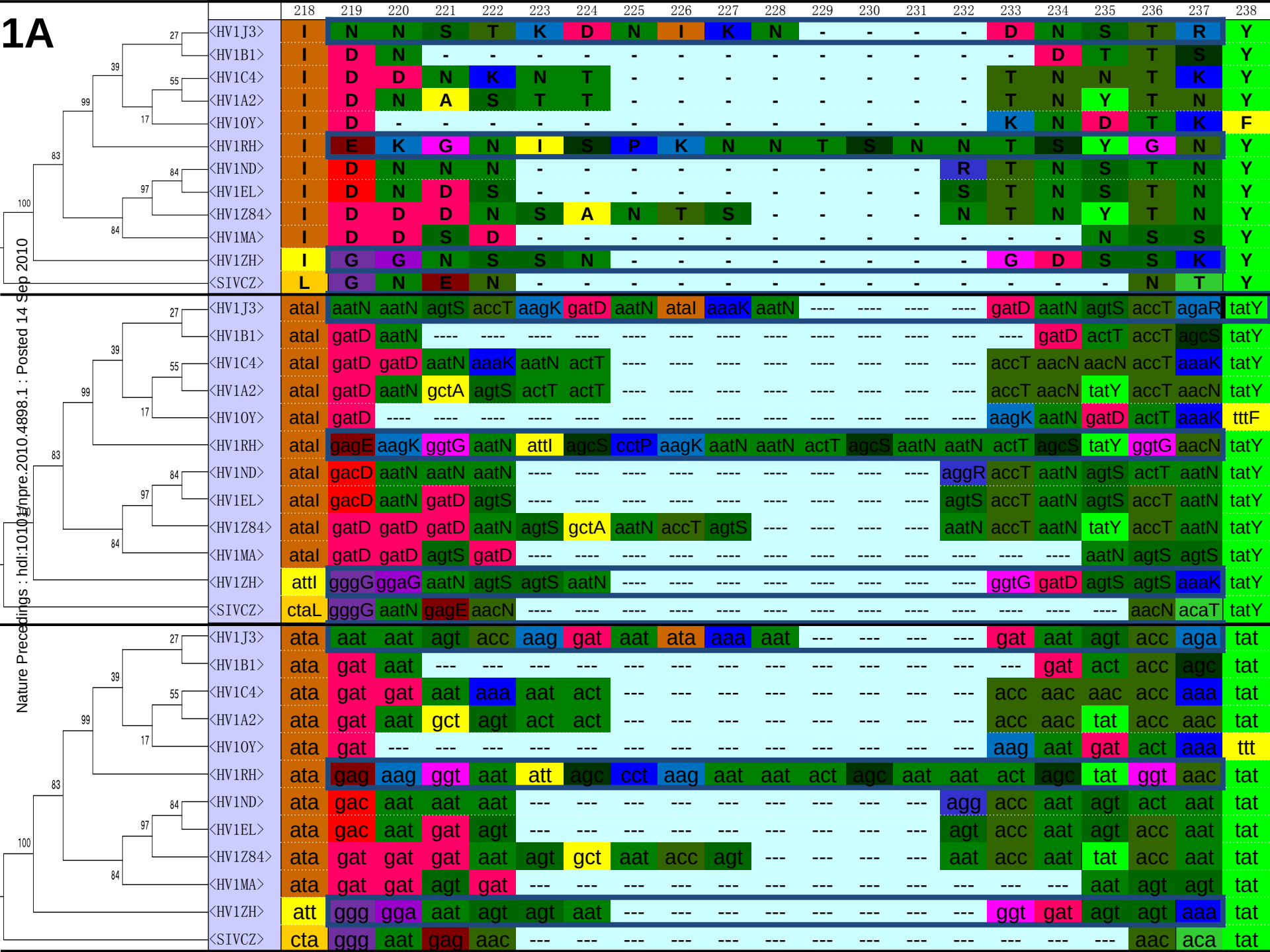
Fig. 1. Comparison of combined and protein-only sequence alignments of HIV gp120. **(A)** A protein-only alignment aligned by ClustalW is reverse translated and visualized in protein-view (top), combined-view (middle), and DNA-view (bottom). **(B)** A combined sequence alignment aligned by DNA^{+Pro} is visualized in combined-view (top), protein-view (middle) and DNA-view (bottom). In the combined sequences, every triplet codon is immediately followed by the one-letter code for its encoded amino acid. DNA and protein sequences are written respectively in lowercase and uppercase letters. The phylogenetic tree shown in the left are inferred from the corresponding alignment. Three representative inserted sequences are shown by blue boxes, and their evolutionary relationships are shown by blue arrows.

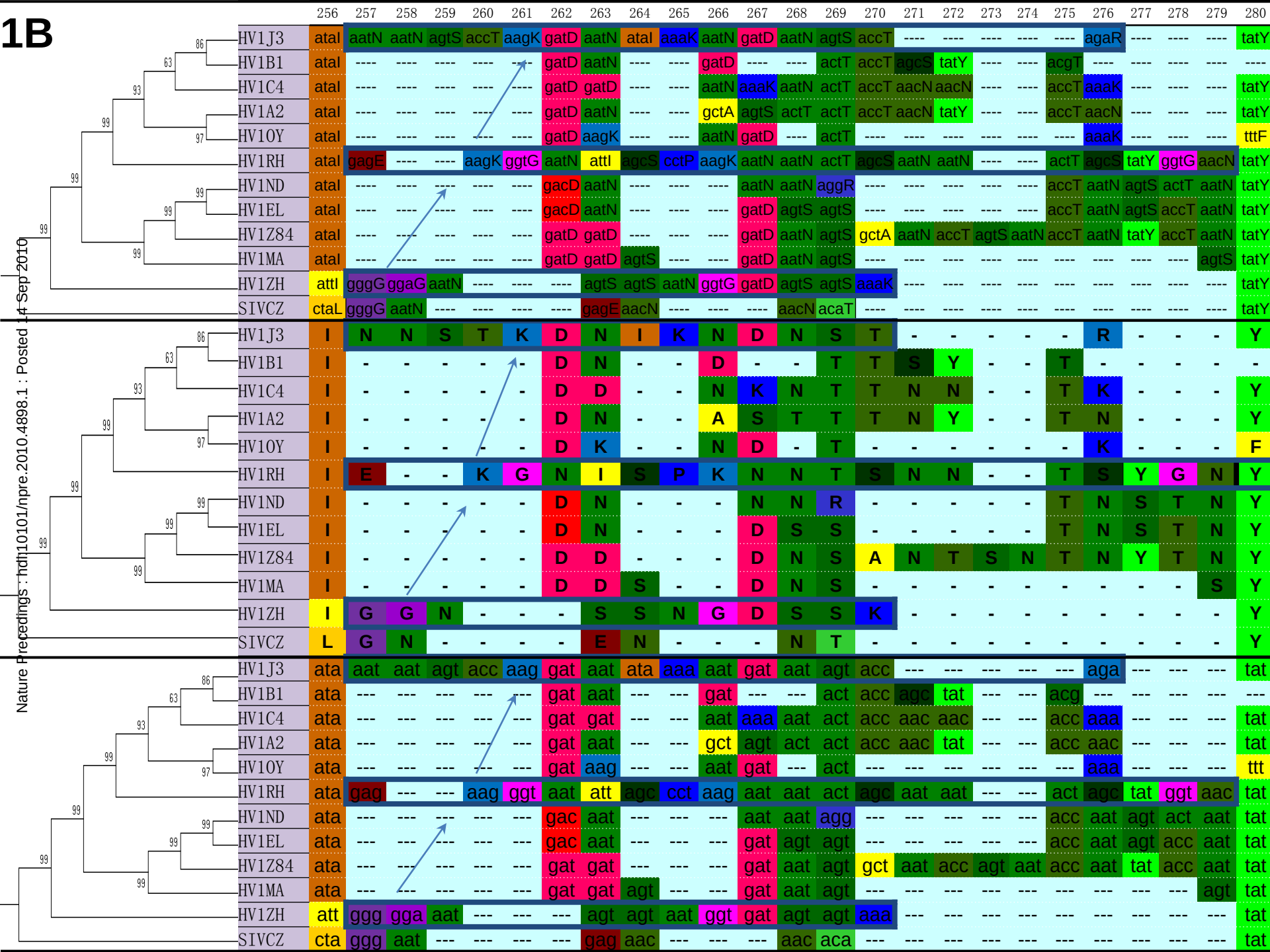
Fig. 2. Combined DNA-protein alignments suggest consistent evolutionary process for different HIV genes. **(A)** The phylogenetic tree for *env* and *gag* genes constructed from protein-only alignments are inconsistent. **(B)** The phylogenetic tree inferred from combined DNA-protein alignments suggests a consistent evolutionary process both for *env* and *gag*.

Fig. 3. Combined alignments suggest consistent evolutionary tree for restriction enzyme Sausa300_2431 homologs. **(A)** The protein-only phylogenetic tree for Sausa300_2431 homologs. **(B)** The combined phylogenetic tree for Sausa300_2431 homologs (C) The robust multi-gene phylogenetic tree for *Staphylococcus aureus subsp. aureus*.

Fig. S1. Comparison of gp120 sequence alignments aligned by different algorithms. **(A)** ClustalW**, (B)** MAFFT, **(C)** MUSSLE, **(D)** T-coffee, **(E)** PRANK, **(F)** DNA^{+Pro} . The phylogenetic trees (show in the left) inferred from different alignments are all inconsistent, only the phylogenetic tree from the combined alignment suggests a clear evolutionary process for the inserted sequences.

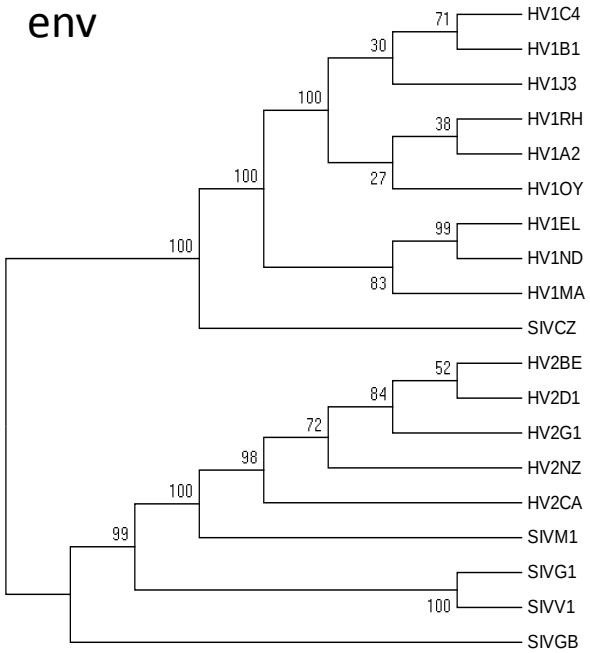
Fig. S2. Comparison of phylogenetic tree constructed from alignments aligned by different algorithms. **(A)** ClustalW**, (B)** MAFFT, **(C)** MUSSLE, **(D)** T-coffee, **(E)** PRANK, **(F)** DNA^{+Pro} .

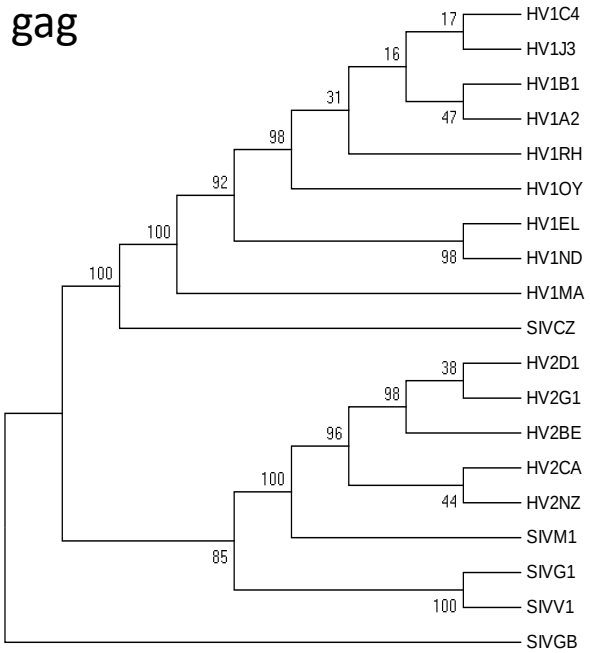



1B
256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280
HV1J3 atal aatN aatN agtS accT aagK gatD aatN atal aaaK aatN gatD aatN agtS accT ---- ---- ---- ---- ---- agaR ---- ---- ---- tatY
HV1B1 atal ---- ---- ---- ---- ---- gatD aatN ---- ---- gatD ---- ---- actT accT agcS tatY ---- ---- acgT ---- ---- ---- ---- ----
HV1C4 atal ---- ---- ---- ---- ---- gatD gatD ---- ---- aatN aaaK aatN actT accT aacN aacN ---- ---- accT aaaK ---- ---- ---- tatY
HV1A2 atal ---- ---- ---- ---- ---- gatD aatN ---- ---- gctA agtS actT actT accT aacN tatY ---- ---- accT aacN ---- ---- ---- tatY
HV1OY atal ---- ---- ---- ---- ---- gatD aagK ---- ---- aatN gatD ---- actT ---- ---- ---- ---- ---- ---- aaaK ---- ---- ---- tttF
HV1RH atal gagE ---- ---- aagK ggtG aatN attI agcS cctP aagK aatN aatN actT agcS aatN aatN ---- ---- actT agcS tatY ggtG aacN tatY
HV1ND atal ---- ---- ---- ---- ---- gacD aatN ---- ---- ---- aatN aatN aggR ---- ---- ---- ---- ---- accT aatN agtS actT aatN tatY
HV1EL atal ---- ---- ---- ---- ---- gacD aatN ---- ---- ---- gatD agtS agtS ---- ---- ---- ---- ---- accT aatN agtS accT aatN tatY
HV1Z84 atal ---- ---- ---- ---- ---- gatD gatD ---- ---- ---- gatD aatN agtS gctA aatN accT agtS aatN accT aatN tatY accT aatN tatY
HV1MA atal ---- ---- ---- ---- ---- gatD gatD agtS ---- ---- gatD aatN agtS ---- ---- ---- ---- ---- ---- ---- ---- ---- agtS tatY
HV1ZH attI gggG ggaG aatN ---- ---- ---- agtS agtS aatN ggtG gatD agtS agtS aaaK ---- ---- ---- ---- ---- ---- ---- ---- ---- tatY
SIVCZ ctaL gggG aatN ---- ---- ---- ---- gagE aacN ---- ---- ---- aacN acaT ---- ---- ---- ---- ---- ---- ---- ---- ---- tatY
HV1J3 I N N S T K D N I K N D N S T - - - - - R - - - Y
HV1B1 I - - - - - D N - - D - - T T S Y - - T - - - - -
HV1C4 I - - - - - D D - - N K N T T N N - - T K - - - Y
HV1A2 I - - - - - D N - - A S T T T N Y - - T N - - - Y
HV1OY I - - - - - D K - - N D - T - - - - - - K - - - F
HV1RH I E - - K G N I S P K N N T S N N - - T S Y G N Y
HV1ND I - - - - - D N - - - N N R - - - - - T N S T N Y
HV1EL I - - - - - D N - - - D S S - - - - - T N S T N Y
HV1Z84 I - - - - - D D - - - D N S A N T S N T N Y T N Y
HV1MA I - - - - - D D S - - D N S - - - - - - - - - S Y
HV1ZH I G G N - - - S S N G D S S K - - - - - - - - - Y
SIVCZ L G N - - - - E N - - - N T - - - - - - - - - Y
HV1J3 ata aat aat agt acc aag gat aat ata aaa aat gat aat agt acc --- --- --- --- --- aga --- --- --- tat
HV1B1 ata --- --- --- --- --- gat aat --- --- gat --- --- act acc agc tat --- --- acg --- --- --- --- ---
HV1C4 ata --- --- --- --- --- gat gat --- --- aat aaa aat act acc aac aac --- --- acc aaa --- --- --- tat
HV1A2 ata --- --- --- --- --- gat aat --- --- gct agt act act acc aac tat --- --- acc aac --- --- --- tat
HV1OY ata --- --- --- --- --- gat aag --- --- aat gat --- act --- --- --- --- --- --- aaa --- --- --- ttt
HV1RH ata gag --- --- aag ggt aat att agc cct aag aat aat act agc aat aat --- --- act agc tat ggt aac tat
HV1ND ata --- --- --- --- --- gac aat --- --- --- aat aat agg --- --- --- --- --- acc aat agt act aat tat
HV1EL ata --- --- --- --- --- gac aat --- --- --- gat agt agt --- --- --- --- --- acc aat agt acc aat tat
HV1Z84 ata --- --- --- --- --- gat gat --- --- --- gat aat agt gct aat acc agt aat acc aat tat acc aat tat
HV1MA ata --- --- --- --- --- gat gat agt --- --- gat aat agt --- --- --- --- --- --- --- --- --- agt tat
HV1ZH att ggg gga aat --- --- --- agt agt aat ggt gat agt agt aaa --- --- --- --- --- --- --- --- --- tat
SIVCZ cta ggg aat --- --- --- --- gag aac --- --- --- aac aca --- --- --- --- --- --- --- --- --- tat
Nature Precedings : hdl:10101/npre.2010.4898.1 : Posted 14 Sep 2010

2A

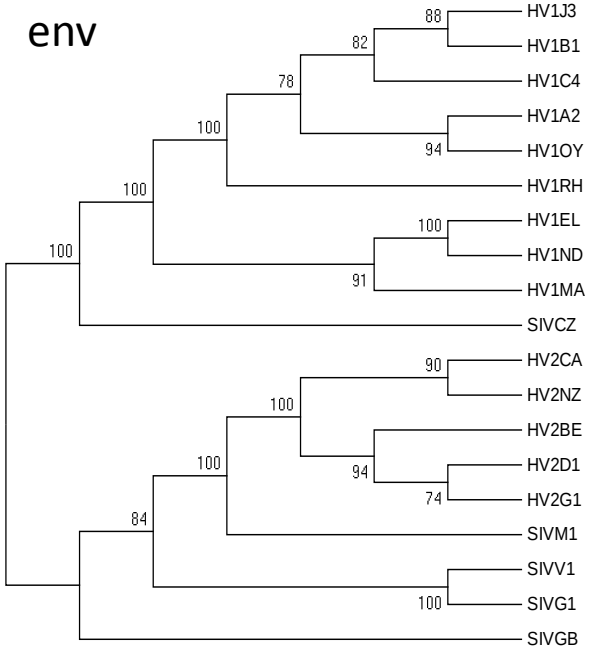
Protein-only

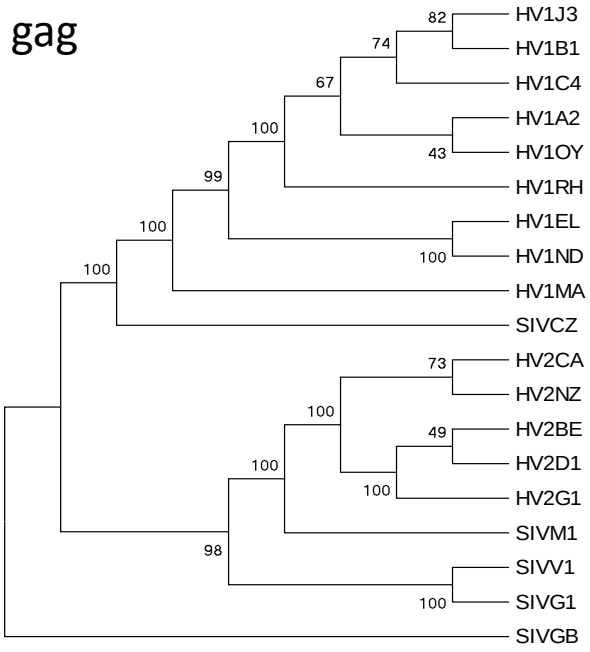




2B

DNA^{+PRO}

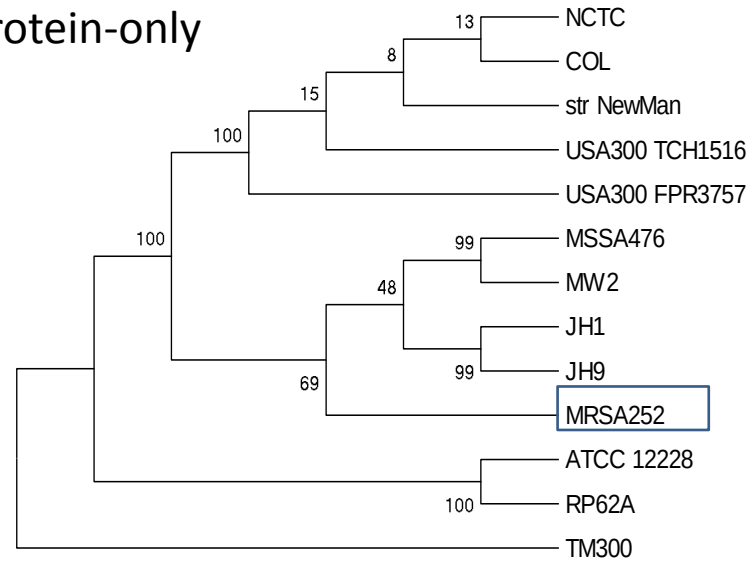




SAUSA300_2431 homologs

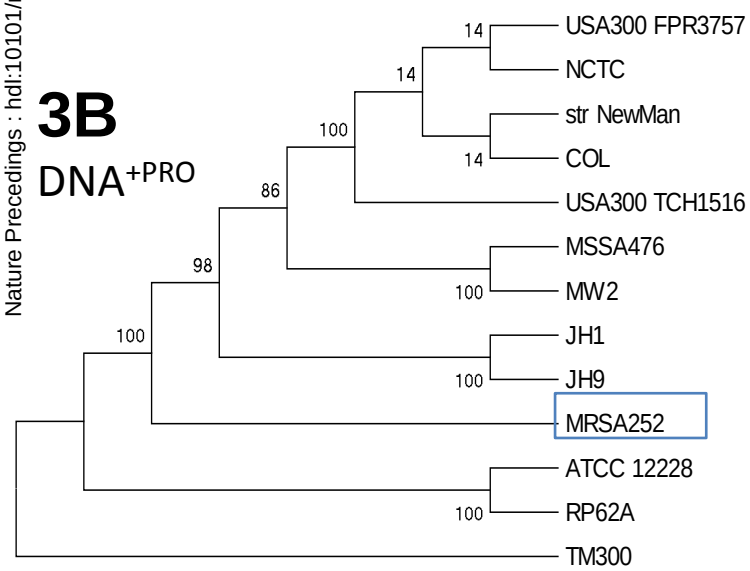
3A

Protein-only



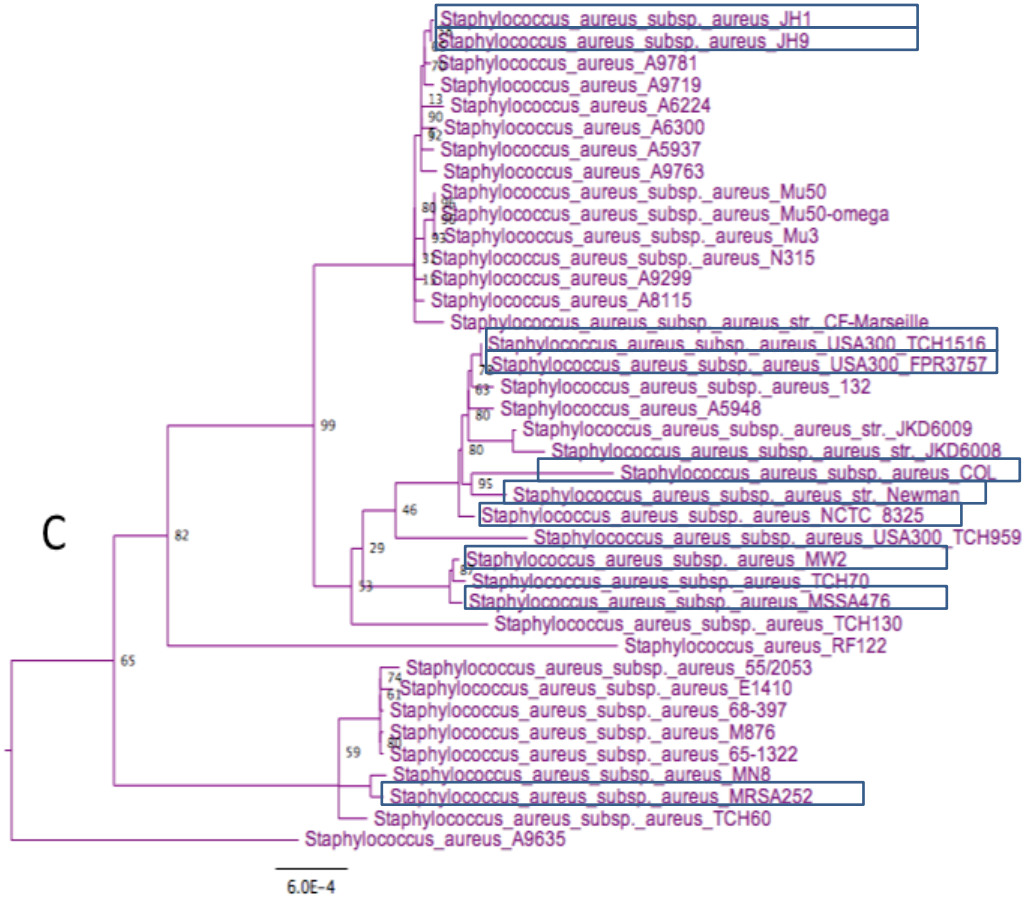
3B

DNA^{+PRO}



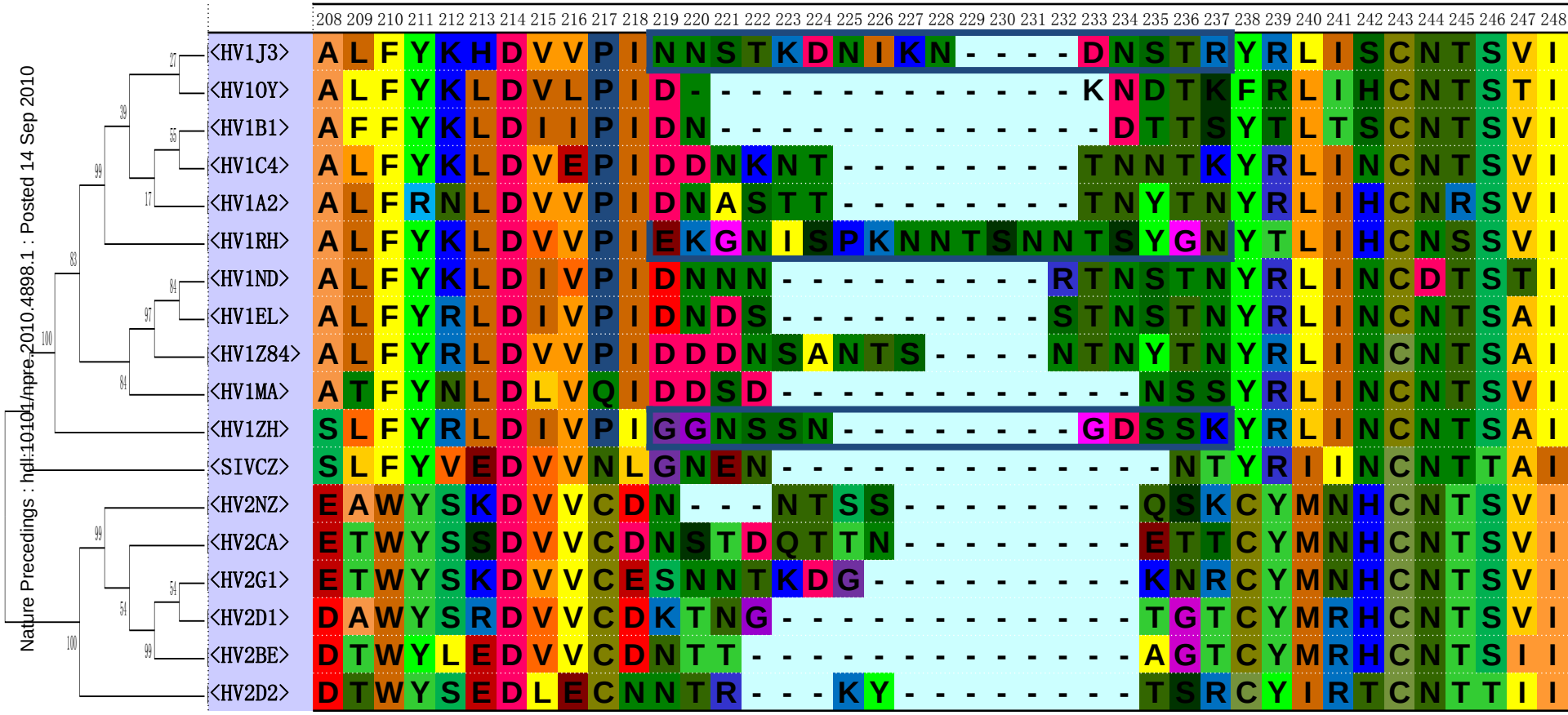
3C

A Robust multi-gene phylogenetic tree



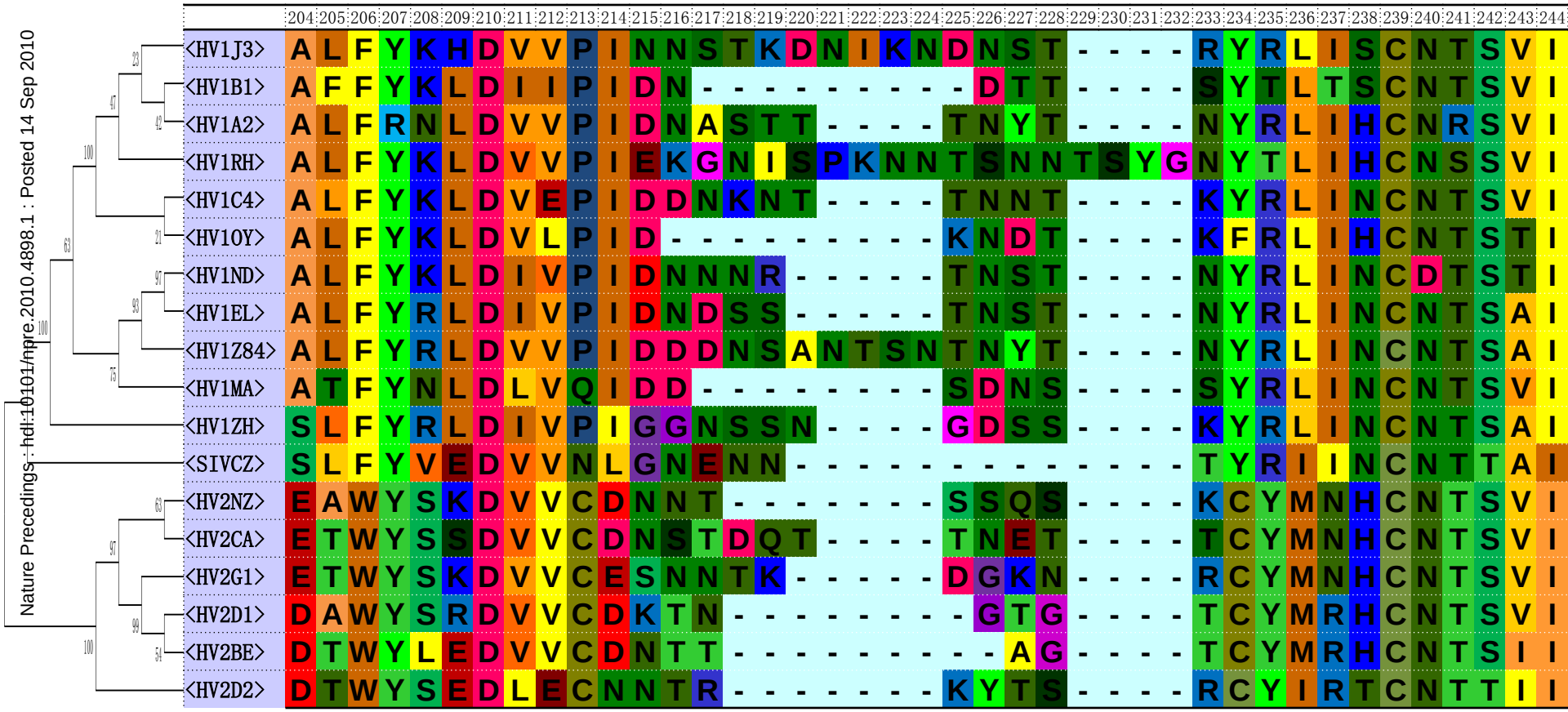
S1A

ClustalW



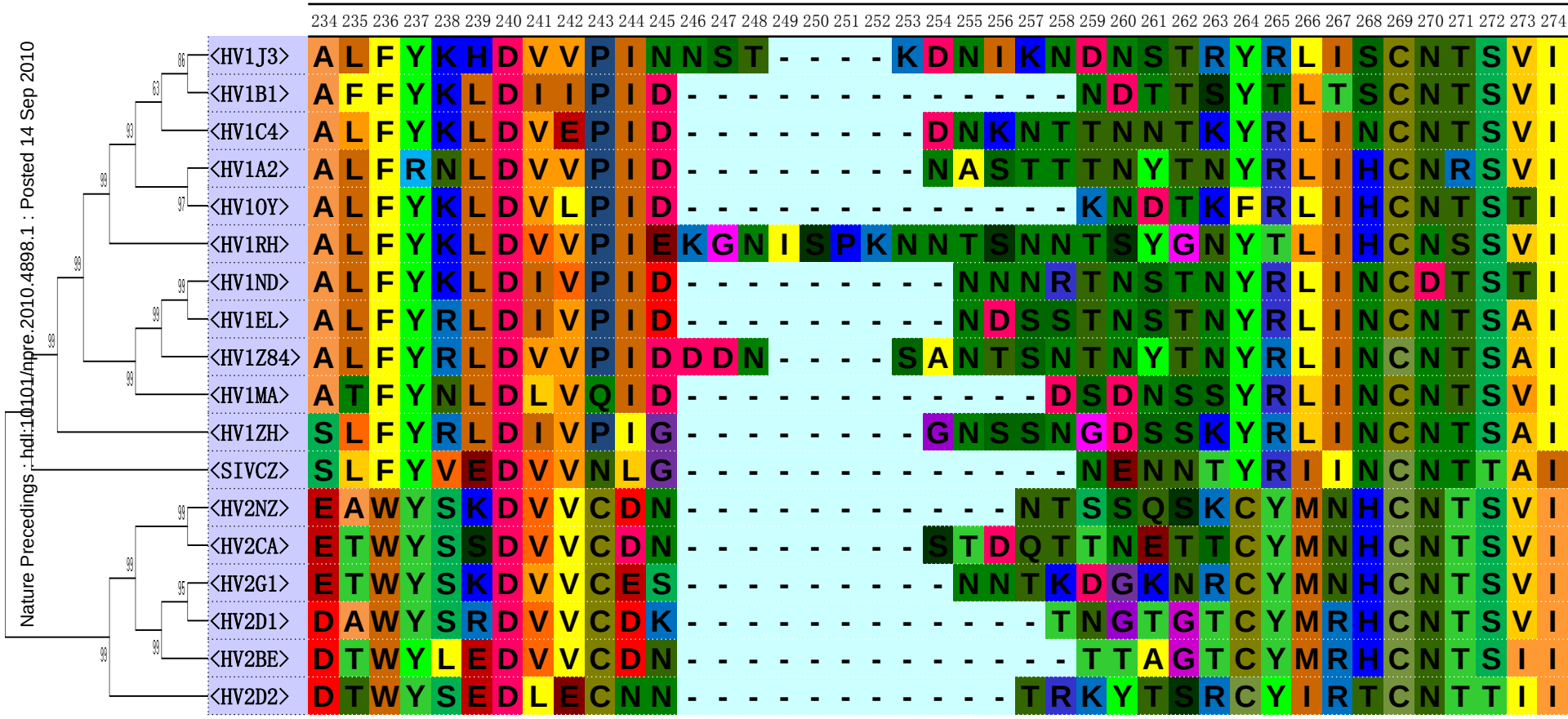
S1B

MAFFT



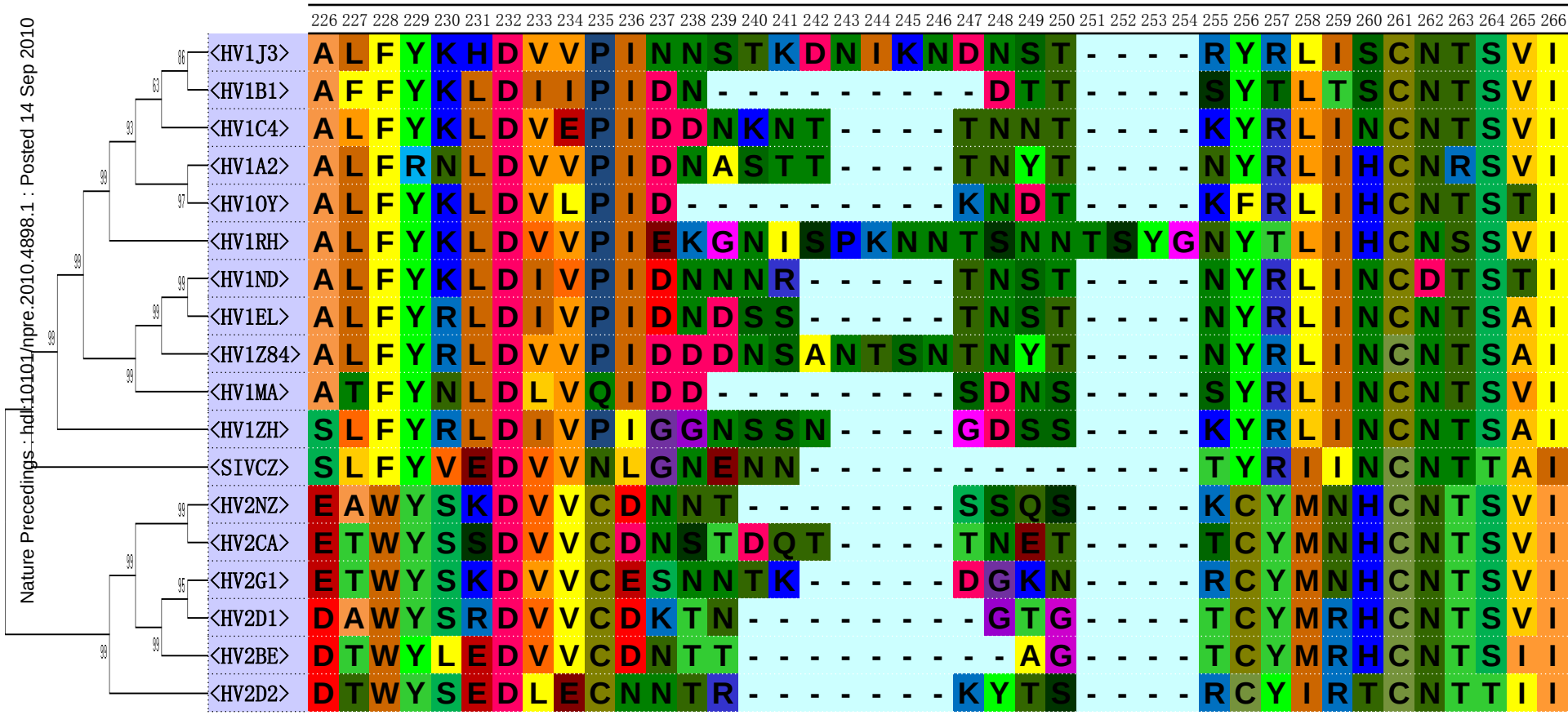
S1C

MUSSLE



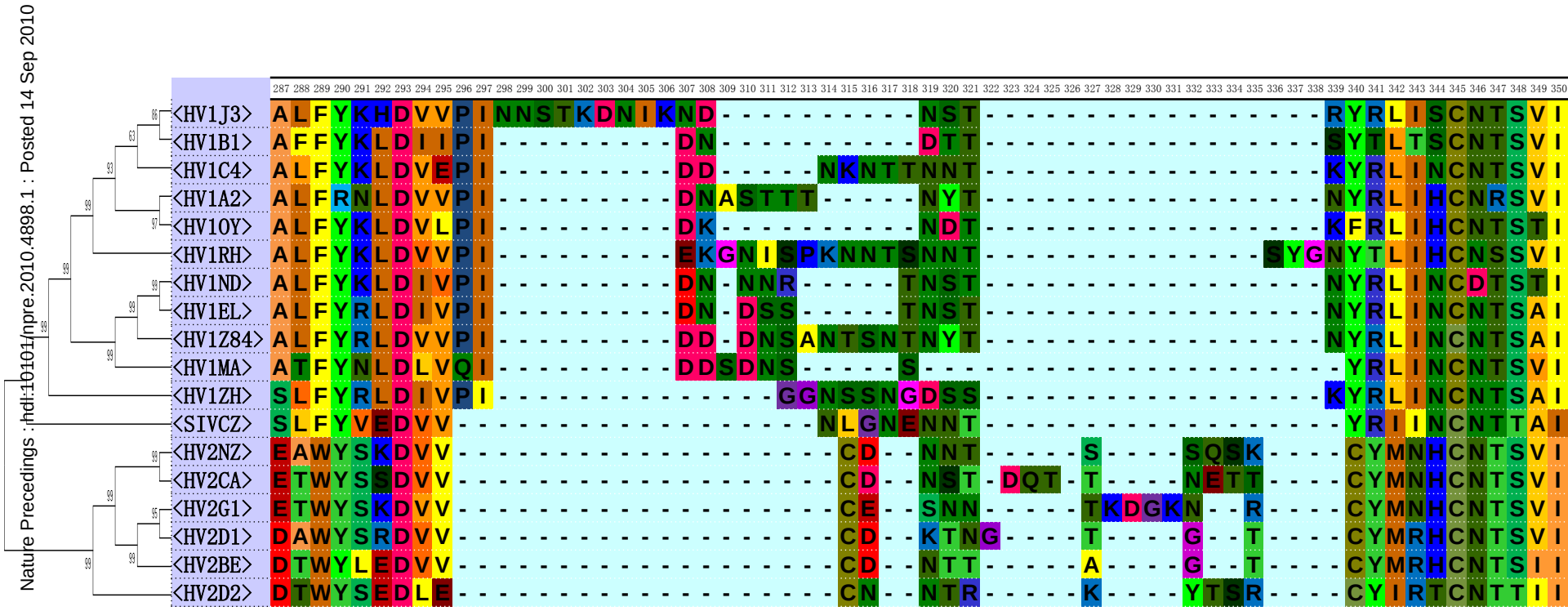
S1D

T-coffee


Nature Precedings : hdl:10101/npre.2010.4898.1 : Posted 14 Sep 2010
226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266
<HV1J3>
<HV1B1>
<HV1C4>
<HV1A2>
<HV1OY>
<HV1RH>
<HV1ND>
<HV1EL>
<HV1Z84>
<HV1MA>
<HV1ZH>
<SIVCZ>
<HV2NZ>
<HV2CA>
<HV2G1>
<HV2D1>
<HV2BE>
<HV2D2>

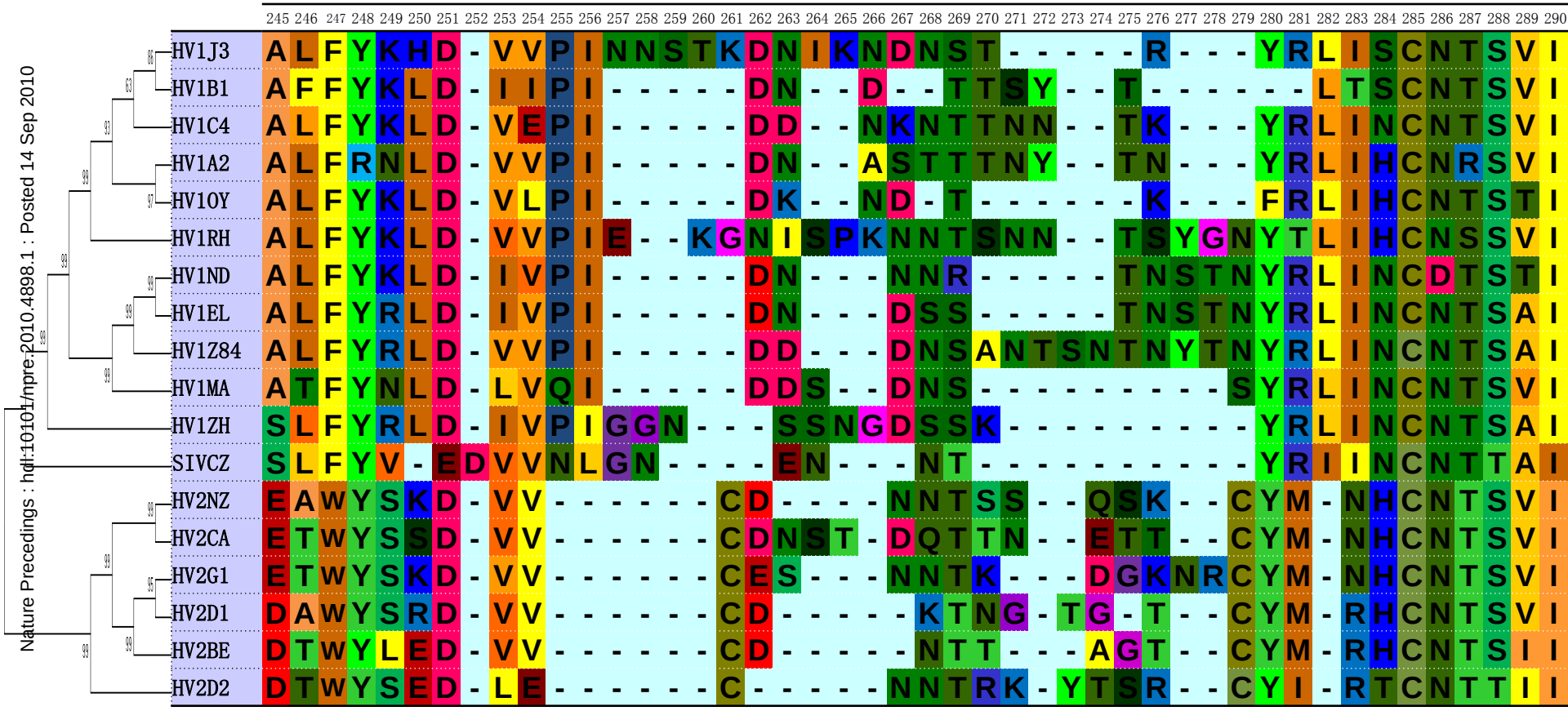
S1E

PRANK


Nature Precedings : hdl:10101/npre.2010.4898.1 : Posted 14 Sep 2010
287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350
<HV1J3>
<HV1B1>
<HV1C4>
<HV1A2>
<HV1OY>
<HV1RH>
<HV1ND>
<HV1EL>
<HV1Z84>
<HV1MA>
<HV1ZH>
<SIVCZ>
<HV2NZ>
<HV2CA>
<HV2G1>
<HV2D1>
<HV2BE>
<HV2D2>

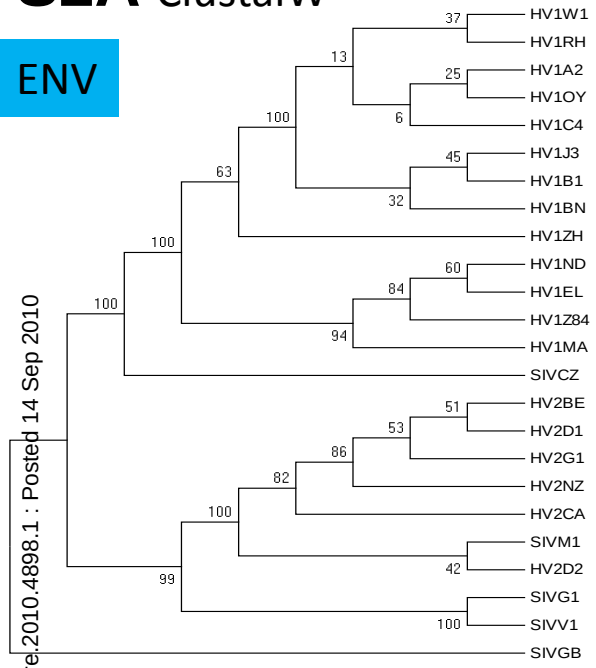
S1F

DNA^{+PRO}

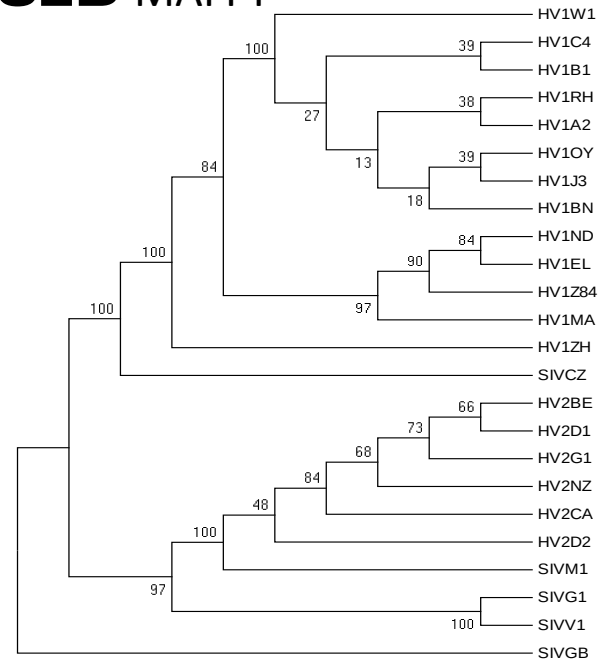


ENV

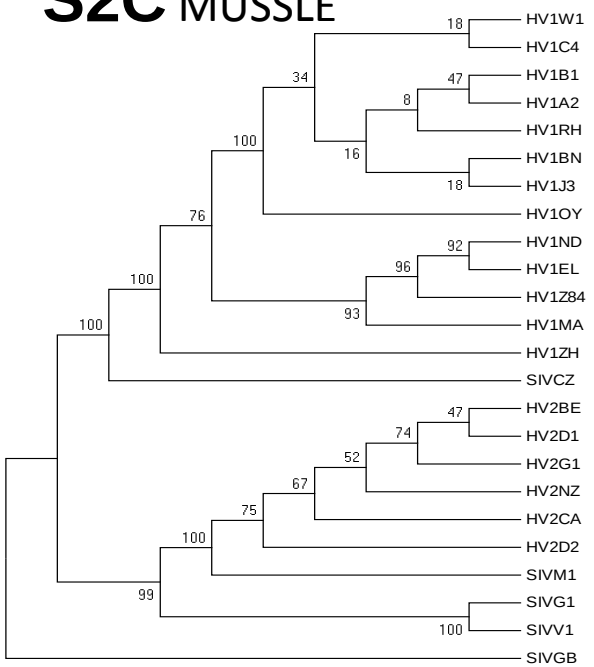
S2A ClustalW



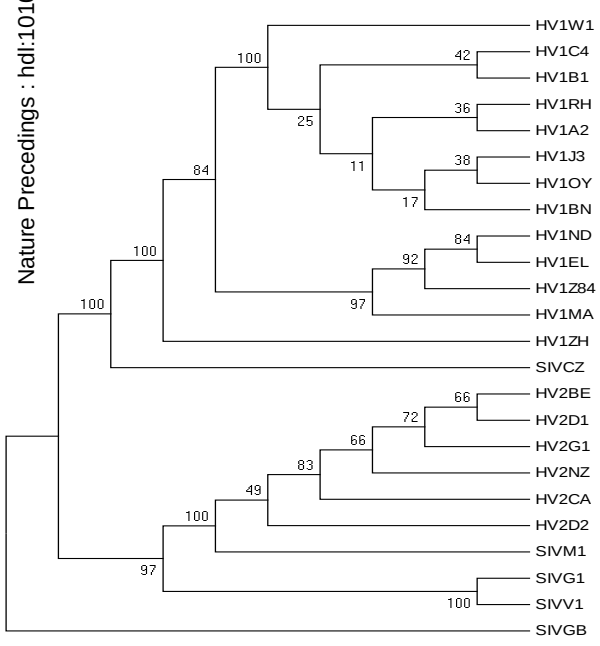
S2B MAFFT



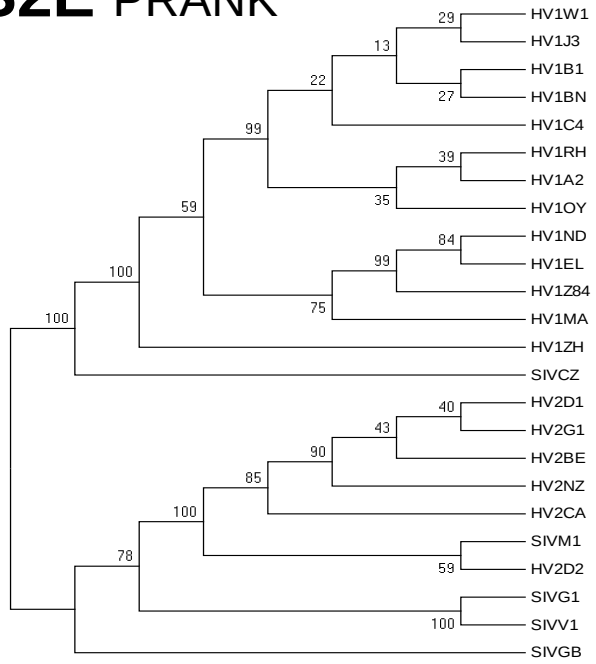
S2C MUSSLE



S2D T-coffee



S2E PRANK



S2F DNA+PRO

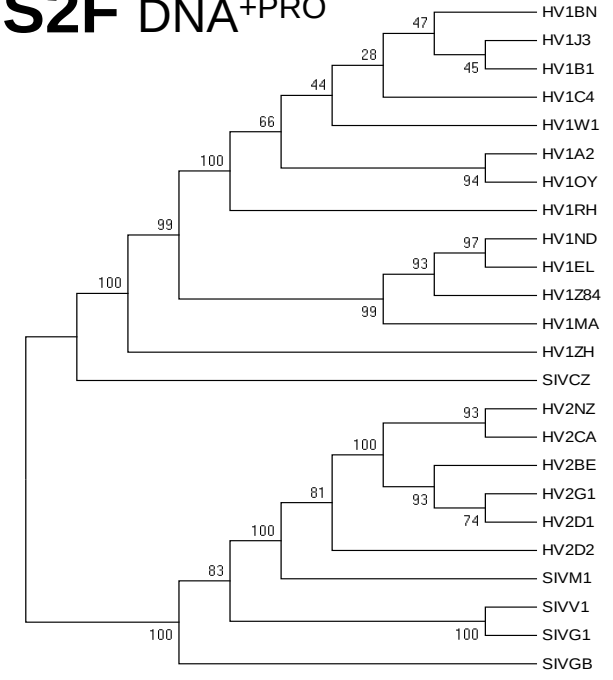


Table 1. The CDP-Gon250 scoring matrix.

	a	c	g	u	B (A)	X (C)	D	E	F	Z (G)	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
a	20	-10	-10	-10	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	a
c	-10	20	-10	-10	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	c
G	-10	-10	20	-10	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	G
u	-10	-10	-10	20	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99	u
B(A)	-99	-99	-99	-99	24	5	-3	0	-23	5	-8	-8	-4	-12	-7	-3	3	-2	-6	11	6	1	-36	-22	B(A)
X(C)	-99	-99	-99	-99	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	X(C)
D	-99	-99	-99	-99	-3	-32	47	27	-45	1	4	-38	5	-40	-30	22	-7	9	-3	5	0	-29	-52	-28	D
E	-99	-99	-99	-99	0	-30	27	36	-39	-8	4	-27	12	-28	-20	9	-5	17	4	2	-1	-19	-43	-27	E
F	-99	-99	-99	-99	-23	-8	-45	-39	70	-52	-1	10	-33	20	16	-31	-38	-26	-32	-28	-22	1	36	51	F
Z(G)	-99	-99	-99	-99	5	-20	1	-8	-52	66	-14	-45	-11	-44	-35	4	-16	-10	-10	4	-11	-33	-40	-40	Z(G)
H	-99	-99	-99	-99	-8	-13	4	4	-1	-14	60	-22	6	-19	-13	12	-11	12	6	-2	-3	-20	-8	22	H
I	-99	-99	-99	-99	-8	-11	-38	-27	10	-45	-22	40	-21	28	25	-28	-26	-19	-24	-18	-6	31	-18	-7	I
K	-99	-99	-99	-99	-4	-28	5	12	-33	-11	6	-21	32	-21	-14	8	-6	15	27	1	1	-17	-35	-21	K
L	-99	-99	-99	-99	-12	-15	-40	-28	20	-44	-19	28	-21	40	28	-30	-23	-16	-22	-21	-13	18	-7	0	L
M	-99	-99	-99	-99	-7	-9	-30	-20	16	-35	-13	25	-14	28	43	-22	-24	-10	-17	-14	-6	16	-10	-2	M
N	-99	-99	-99	-99	-3	-18	22	9	-31	4	12	-28	8	-30	-22	38	-9	7	3	9	5	-22	-36	-14	N
P	-99	-99	-99	-99	3	-31	-7	-5	-38	-16	-11	-26	-6	-23	-24	-9	76	-2	-9	4	1	-18	-50	-31	P
Q	-99	-99	-99	-99	-2	-24	9	17	-26	-10	12	-19	15	-16	-10	7	-2	27	15	2	0	-15	-27	-17	Q
R	-99	-99	-99	-99	-6	-22	-3	4	-32	-10	6	-24	27	-22	-17	3	-9	15	47	-2	-2	-20	-16	-18	R
S	-99	-99	-99	-99	11	1	5	2	-28	4	-2	-18	1	-21	-14	9	4	2	-2	22	15	-10	-33	-19	S
T	-99	-99	-99	-99	6	-5	0	-1	-22	-11	-3	-6	1	-13	-6	5	1	0	-2	15	25	0	-35	-19	T
V	-99	-99	-99	-99	1	0	-29	-19	1	-33	-20	31	-17	18	16	-22	-18	-15	-20	-10	0	34	-26	-11	V
W	-99	-99	-99	-99	-36	-10	-52	-43	36	-40	-8	-18	-35	-7	-10	-36	-50	-27	-16	-33	-35	-26	142	41	W
Y	-99	-99	-99	-99	-22	-5	-28	-27	51	-40	22	-7	-21	0	-2	-14	-31	-17	-18	-19	-19	-11	41	78	Y
	a	c	g	u	B(A)	X(C)	D	E	F	Z(G)	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

Table 2. Optimal parameters for Clustal W to align combined DNA-Protein sequences

Option	**Parameter**	**Description**
-INFILE	Input.fasta	Input merged DNA-Protein sequence in FASTA format
-TYPE	PROTEIN	protein alignment
-NEGATIVE	ON	protein alignment with negative values in matrix
-OUTFILE	Output.fasta	Output sequence alignment file name
-OUTPUT	PIR	Output alignment file in FASTA format
-PWMATRIX	DNA10-3-Gon250.txt	User-defined pairwise alignments scoring matrix (Table 1)
-PWGAPOPEN	10	Pairwise alignments gap opening penalty
-PWGAPEXT	0.1~0.2	Pairwise alignments gap opening penalty
-MATRIX	DNA3-1-Blossum62.txt	User-defined multiple alignments scoring matrix (Table 2)
-GAPOPEN	3	Multiple alignments gap opening penalty
-GAPEXT	02~0.5	Multiple alignments gap opening penalty
-ENDGAPS	NO	No end gap separation pen.
-GAPDIST	4	Gap separation pen. range
-NOPGAP		Residue-specific gaps off
-NOHGAP		Hydrophilic gaps off
-ITERATION	NONE	Perform iteration at each step to improve the alignment.

REFERENCES

1. Thompson J.D., Higgins, D.G. and Gibson, T.J. (1994) **CLUSTAL W:** improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673–4680.
2. Löytynoja A, Goldman N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* 320 (5883):1632-5.
3. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, Hahn BH, Sharp PM. (2003) Hybrid origin of SIV in chimpanzees. *Science*. 300 (5626):1713.
4. Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan. (1995) Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res. Hum. Retrovir.* 11:1423–1425.
5. Anderson JP, Rodrigo AG, Learn GH, Madan A, Delahunty C, Coon M, Girard M, Osmanov S, Hood L, Mullins JI. (2000) Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. *J Virol*. 74(22):10752-65.
6. Wernersson R, Pedersen AG. (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* 31(13): 3537-9.
7. Hein J. (1994) An algorithm combining DNA and protein alignment. *J. Theor. Biol.*, 167, 169–174.
8. Hein J. and Støvlbæk, J. (1996) Combined DNA and protein alignment. *Methods Enzymol.*, 266, 402–418.
9. Robert C Edgar, 2004; **MUSCLE**: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 5: 113.
10. Robert C. Edgar, 2004; **MUSCLE**: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5): 1792–1797.
11. Timo Lassmann and Erik LL Sonnhammer, 2005; **Kalign** – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics.* 6: 298.
12. Iain M. Wallace, Orla O'Sullivan, Desmond G. Higgins, and Cedric Notredame, 2006; **M-Coffee**: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. 34(6): 1692–1699.
13. V. A. Simossis and J. Heringa, 2005; **PRALINE**: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.* 33.
14. Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata, 2005; **MAFFT** version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33(2): 511–518.
15. Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata, 2002; **MAFFT**: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30(14): 3059–3066.
16. Iain M. Wallace, Orla O'Sullivan, Desmond G. Higgins, and Cedric Notredame, 2006; **M-Coffee:** combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* 34(6): 1692–1699.
17. Morgenstern B. (1999) **DIALIGN** 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15, 211–218.

18. Chuong B. Do, Mahathi S.P. Mahabhashyam, Michael Brudno, and Serafim Batzoglou, 2005; **ProbCons**: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* 15(2): 330–340.
19. Ari Löytynoja and Nick Goldman, 2005; An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA.* 102(30): 10557–10562.
20. Cédric Notredame, 2007; Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Comput Biol.* 3(8): e123.
21. Snyder EE, Kampanya N, Lu J, *et al*. 2007; **PATRIC**: The VBI PathoSystems Resource Integration Center. *Nucleic Acids Res.* 35 (Database issue) 401-406.