CAUSA 2.0: accurate and consistent evolutionary analysis of proteins using codon and amino acid unified sequence alignments

Xiaolong Wang, Chao Yang

Multiple sequence alignment (MSA) is widely used to reveal structural and functional changes leading to genetic differences among species, and to reconstruct evolutionary histories of related genes, proteins and genomes. Traditionally, proteins and their coding sequences (CDSs) are aligned and analyzed separately, but often drastically different conclusions were drawn on a same set of data. Here we present a new alignment strategy, *Codon and Amino Acid Unified Sequence Alignment (CAUSA) 2.0*, which aligns proteins and their coding sequences simultaneously. CAUSA 2.0 optimizes the alignment of CDSs at both codon and amino acid level efficiently. Theoretical analysis showed that CAUSA 2.0 enhances the entropy information content of MSA. Empirical data analysis demonstrated that CAUSA 2.0 is more accurate and consistent than nucleotide, protein or codon level alignments. CAUSA 2.0 locates in-frame indels more accurately, makes the alignment of coding sequences biologically more significant, and reveals several novel mutation mechanisms that relate to some genetic diseases. CAUSA 2.0 is available in website www.DNAPlusPro.com .

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1214v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

CAUSA 2.0: accurate and consistent evolutionary analysis of proteins using codon and amino acid unified sequence alignments

Xiaolong Wang[*], *Chao Yang*

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

*Correspondence:

Xiaolong Wang, Ph.D., Associate Professor,
Department of Biotechnology
Ocean University of China
No. 5 Yushan Road
Qingdao, 266003, Shandong, P. R. China
Tel: 0086-139-6969-3150
E-mail: Xiaolong@ouc.edu.cn

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1214v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

Introduction

In recent years, with increasingly wider availability of sequence data, it has been able to reveal structural and functional changes leading to genetic differences among species, and allows reconstruction of evolutionary histories of related genes, proteins and genomes. These studies require multiple sequence alignment (MSA) for accurate identification of homologous residues and positioning of gaps indicating insertions and deletions (indels). At present, there are quite some MSA tools, such as Clustal W [1,2], MAFFT [3], MUSCLE [4], T-coffee [5] and PRANK [6]. However, different programs often give drastically different alignments on a same set of data, and support entirely different mechanisms driven evolutionary changes. In fact, MSAs are sometimes so inaccurate that they could lead to misinterpretation of evolutionary information [7].

Due to the small size of the alphabet of DNA bases, it is difficult to distinguish true homology from random similarities, and thus the alignment of DNA sequences is inherently inaccurate [8]. Proteins are built from 20 amino acids, so that the ‘signal-to-noise ratio’ in protein sequence alignments is better than that of DNA sequences. Besides this advantage in information-content, protein alignments also benefit from a series of AA substitution matrices, such as PAM [9], BLOSUM [10] and Gonnet [11], which contain empirically derived scores for each possible AA substitution and provide a rational basis for aligning AAs.

Particularly, the problem of aligning the coding sequences (CDSs) for a certain family of proteins has been studied fairly in depth. Various methods have been developed to solve this problem: The first is to construct a codon alignment by back-translating a protein alignment, such as RevTrans [8], PAL2NAL [12] and TranslatorX [13]; The second, Hein’s COMBAT [14,15], is to construct a combined alignment by combining a DNA alignment and a protein alignment; Unfortunately, it has been applied only in pairwise alignment, but not in MSA; The third is to construct a codon alignment by using an empirical codon substitution scoring matrix [16]. In addition, many codon-based models [17-21] has been developed to produce more reliable estimates of certain biologically important measures, such as the transition/transversion rate ratio and the synonymous/nonsynonymous substitution rate ratio, than can nucleotide-based models.

Owing to the complexities of the alignment of CDSs, it has been rather difficult to choose an appropriate tool, method and parameters for MSA and the subsequent evolutionary analysis. Moreover, traditionally protein sequences and their CDSs have been aligned and analyzed separately, in nucleotide, amino acid (AA) or codon level. It is well known that they often give incongruent conclusions, in practice, however, usually only one kind of alignment is selected for the subsequent phylogenetic and evolutionary analysis, while the others are dismissed to avoid the problem of inconsistency in different (nucleotide, AA or codon) alignment methods. Here we present an alternative alignment strategy, *Codon and Amino Acid Unified Sequence Alignment* (CAUSA 2.0), which aligns protein sequences and their CDSs simultaneously. We demonstrated that CAUSA 2.0 is evolutionarily accurate and consistent, computationally efficient, and it reveals some novel mutation mechanisms.

Materials and Methods

Protein coding sequences

Different strains of human and simian immunodeficiency virus were derived from the seed alignment of Pfam family pf00516. The coding sequences (CDSs) of their Envelope glycoprotein gp120 (Env) and core (Gag) proteins were retrieved from the HIV sequence database [22]. Thirty protein families and their standard phylogenetic trees in human and mammalian animals were arbitrarily selected from TreeFam-A (http://www.treefam.org/).

Data analysis

The flowchart of data analysis is shown in Fig 1, DNA or protein alignments were aligned by the multiple sequence alignment tools at EBI (http://www.ebi.ac.uk/), including ClustalW, MAFFT, MUSCLE, T-COFFEE and PRANK. Codon alignments were aligned by PRANK using "align translated codons" option (PRANK-CA), and a codon alignment tool (CAT) provided by the HIV database at the Los Alamos National Laboratory (http://www.hiv.lanl.gov/). All programs were run with their default settings. The principle and implementation of CAUSA 2.0, and the methodologies for the construction of sequence alignments, phylogenetic trees, evolutionary analysis and structure modeling, were described in details in the *Supplementary Material*.

Results

CAUSA 2.0 improves the accuracy of the alignment of CDSs

It has been reported that traditional nucleotide or protein level alignment methods incorrectly squeeze distinct, but similar, inserted sequences between two conserved blocks, thus overestimate the number of substitutions and underestimate that of indels [23]. Using the Pfam seed alignment of Env (PF00516) as a model example, as shown in Fig 2A and Fig S1A, they demonstrated that a traditional ClustalW protein alignment of Env suggested a high rate of substitutions in a variable (V2) region. Alignments of the same region given by MAFFT (Fig S1B), MUSCLE (Fig S1C) and T-coffee (Fig S1D), are basically the same to that of ClustalW. They also pointed out that the problem is caused by repeated penalizing gap-opening [23], but cannot be avoided by reducing gap-opening penalties, since it will result in 'gappy' alignments. In order to solve this problem, they developed PRANK [6,7,23-26], a phylogeny-aware alignment tool, which "flags" gaps introduced in earlier steps, so that distinct insertions are kept separate even when they occur at a same position [23]. As shown in Fig 2B, PRANK identified several distinct insertions.

Using the same set of Env CDSs, we compared CAUSA 2.0 with protein alignments aligned by ClustalW, MAFFT, MUSCLE, T-COFFEE and PRANK, and codon alignments aligned by PRANK-Codon and CAT. As shown in Fig S1A-S1D, all different kinds of protein alignments show a lot of AA substitutions in the V2 region of Env, while in PRANK (Fig 2B, S1E) and PRANK-Codon (Fig 2C, S1F), they are identified as distinct insertions, and thus, many more gaps were inserted.

In CAUSA 2.0 non-synonymous substitutions (NSSs) are displayed in obviously different colors; and synonymous substitutions (SSs) are shown in similar but slightly different colors, makes it much easier for a user to distinguish an SS from a NSS. As indicated by the red boxes in Fig 2, in all of the other alignments, a row of highly conserved Y residues were misaligned. While in CAUSA 2.0 the alignment of this row is obviously more accurate. It is more clearly shown in Fig S1H, the CAUSA 2.0 alignment is visually highly ordered, since the number of substitutions was minimized both in codon and AA level. Similar results could not be obtained with the other MSA by changing their gap penalties. Alignments created by the other programs often require fine adjustments made by human visual inspection before phylogeny analysis, which is cumbersome and a potential source of errors. Using CAUSA 2.0, the need of visual control or manual editing of MSAs is eliminated.

CAUSA 2.0 better interprets molecular evolution

Frequently, CAUSA 2.0 inserts one or more gaps within a codon-AA 4-tuple. In the previous version (CAUSA 1.0), we took them as errors, and had to fix the alignments a posteriori by moving the gaps inserted out of the tuple [27]. However, later we found that gap-moving is totally unnecessary. In the present version (CAUSA 2.0), the gapped codon-AA tuples are kept and called *split codons*. Actually, the split codons are very informative and useful, as they allow CAUSA 2.0 locates in-frame indels more accurately, and reveals some novel mutation mechanisms.

An in-frame indel is an indel in a CDS with a length of a multiple of 3bp, and thus, maintained the open reading frame (ORF). It has been reported that more than half of the functionally important coding indels in human genome were in-frame, but they often do not coincide perfectly with codon boundaries [28]. However, in a traditional protein or codon alignment, every codon is indivisible, and thus, an in-frame indel has to be forced to start from the first base of a codon, leads to an inaccurate alignment of the CDSs involved. By allowing the split codons, CAUSA 2.0 locates in-frame and non-in-frame indels more accurately, and reveals several new mutation mechanisms by which new codons could be generated:

(1) *Codon splitting*: when an in-frame insertion was inserted into an internal location (between 1-2 or 2-3) of a codon, the codon was split into two parts, and fused with the bases in the two ends of the inserted sequence, respectively, forming two new codons.

(2) *Codon fusion*: when an in-frame deletion started from the second or third base of a codon, the remaining one or two bases of this codon fused with the bases in the other side of the deletion, forming a new codon.

(3) *Substitution induced deletion* (*siDel*): when a non-synonymous substitution occurred in a codon, if the encoded AA became similar to another AA nearby in the same sequence, it may allow the deletion of a fragment between them.

(4) *indel-induced partial frame shift* (*ipFS*): when two or more non-in-frame indels neighbored, there will be one or more partial frame shifts occurred within or nearby the indels, if it does not change the whole reading frame of the CDSs.

For an example of codon splitting, as shown in Fig 3A and 3B, ClustalW suggested a lot of independent substitutions between the two groups of virus (HIV1 and HIV2/SIV), PRANK-Codon shows only a few indels and substitutions. However, as shown in Fig 3C, CAUSA 2.0 split codon 670 (Ggg-g) and located the 6-nt insertion (a gca ct) between the 2nd and the 3rd base of this codon. Comparing the three alignments, CAUSA 2.0 might have the highest probability of correctness, since it has the highest number of matches, and requires the least number of mutation events.

In all HIV strains, HIV-1 subtypes are the most common form, they are the most infectious and pathogenic to human, and among the most genetically diverse. During viral penetration into host cell, the fusion peptide region involves in the merging of the virus envelop with host endosomal membrane. The inserted motif is highly conserved in all HIV-1 strains, and is the only difference located within the fusion peptide region compared with the other strains. Therefore, we postulate that the insertion might be a key event lead to the origin of HIV-1, and a critical motif in the interaction of the virus with human cell during the infection process.

For an example of codon fusion, as shown in Fig 3D, ClustalW shows a lot of substitutions, PRANK-Codon shows that there are several indels (Fig 3E). CAUSA 2.0 (Fig 3F), however, identified some smaller, but more accurate, indels. In addition, in different strains the codons between 864 and 867 were deleted and mutated in different ways, however, the amino acid encoded by codon 864 (V or C), is highly conserved among all strains, so it might probably be functionally important.

Note that although a codon spitting could be distinguished easily from a codon fusion according to a phylogenetic tree, it is hard to distinguish an ipFS from a siDel if their ancestral sequence is not available. For example, as shown in Fig 3G and 3H, in ClustalW there are so many base and AA substitutions that PRANK-Codon suggested long indels instead. CAUSA 2.0, however, revealed multiple mutational events, including codon fusion/splitting, siDel and/or ipFS, and gave a different while more accurate alignment for this region (Fig 3I). These evolutionary events bring additional variability, in together with those by the traditional mutation mechanisms, help the HIV to evolve rapidly and develop the ability of drug resistance.

By analyzing the CDSs of thirty protein families in human and mammalian arbitrarily selected from TreeFam-A (Table S3), we found that these new mutation events are not rare, but happened rather frequently not only in virus, but also in human and animals. We postulate that these newly discovered mutation mechanisms may play an important role in the acquisition of new genetic information and provide a potentially powerful way to optimize the protein sequences and structures.

CAUSA 2.0 better explains certain genetic diseases

Over the past decade, millions of small indels have been discovered in human populations and personal genomes [28]. Most of the indels found in protein coding sequences were codon-sized, in-frame indels, and thus maintained the ORF of these proteins. Numerous indels map to functionally important sites within human genes, and are related to human traits and genetic diseases [29-34].

For example, in-frame indels have been identified as one of the main causes for the Wolfram syndrome, an autosomal recessive disorder [35]. Table 1 lists some in-frame indels related to Wolfram syndrome and other genetic diseases, which were deposit in the Human Gene Mutation Database. For example (CD031549), in the coding sequence …K*a*^**ag** A**gca** A**g**^*cc*…, after the deletion of the six nucleotides (in bold), the remaining three nucleotides (in italic) fused into a new codon, Tacc, in which an amino acid substitution occurred without any base substitution. As shown in Fig 4, such kind of disease-related codon splitting/fusion phenomenon is visualized clearly in CAUSA 2.0, but not in the other alignment software, although a nucleotide alignment may sometimes result in the same alignment. Therefore, CAUSA 2.0 would be useful in discovering, recognizing and investigation of this kind of genetic diseases

CAUSA 2.0 improves molecular phylogenetic analyses in virus

For several reasons ranging from methodological issues or bona fide biological phenomena, the phylogenetic trees of different genes for a set of certain genomes are often incongruent topologically. In order to tackle the topological conflicts in different gene trees, phylogenomic studies couple concatenation with practices such as rogue taxon removal, or the use of slowly evolving genes. Recently, however, Salichos and Rokas questioned the exclusive reliance on concatenation and associated practices. In a phylogenomic analysis of 1,070 orthologues from 23 yeast genomes, surprisingly, they identified 1,070 distinct gene trees, which were all incongruent with the phylogeny inferred by concatenation [36].

As shown in Fig S2, when the trees were drawn using neighbor joining (NJ) methods, the phylogenetic trees of Env inferred from ClustalW, MAFFT, T-coffee, MUSCLE and PRANK alignments are all varied, and CAUSA 2.0 suggested another different evolutionary process. Using MSA aligned by ClustalW, when trees are inferred by maximum likelihood (ML) method in “complete deletion” (CD) mode, as shown in Fig 5A-5F, the DNA tree, the back-translate codon tree and the protein tree are slightly inconsistent; however, when the ML trees were drawn in “using all sites” (AS) mode, the AS trees are more inconsistent with each other, and differ greatly with the CD trees (strain HV1MA was clustered in different clade with a very high Bootstrap value), suggesting that in ClustalW the alignment of the variable region is inconsistent with that of the conserved region; In PRANK, as shown in Fig 5G-5L, the DNA tree, the codon tree and the protein tree are more consistent in CD mode, but the CD trees also differ greatly from the AS trees (HV1MA was classified in different clade with a very high Bootstrap value).

In CAUSA 2.0, as shown in Fig 5M-5P, however, the DNA tree and the protein tree are fully consistent in either CD or AS mode, and almost fully consistent between CD and AS mode: the only two inconsistent branches, HV1C4 and HV1A2, stays within a same clade, suggesting that in CAUSA 2.0 the variable region is aligned consistently with the conserved region.

In order to further evaluate the accuracy of different alignments and phylogenetic trees in HIV, we build alignments and trees for two genes, env and gag, respectively using ClustalW, PRANK and CAUSA 2.0. As shown in Fig 6A and 6B, phylogenetic trees for gag protein constructed from ClustalW alignment are inconsistent in different mode, but those from PRANK-Codon (Fig 6C, 6D) and CAUSA 2.0 (Fig 6E, 6F) are all fully consistent with each other in different mode. Moreover, the ClustalW protein trees are different between env (Fig 5A) and gag (Fig 6A), suggesting that in some of these HIV genomes, such as HV1JRFL, HV1J3, HV1B1 and HV1C4, the two genes derived from different strains. However, the PRANK CD trees (Fig 5G, 6C) and CAUSA 2.0 CD trees (Fig 5M, 6E) are basically consistent with each other, and suggest that in most of these virus genomes the two different genes underwent a same evolutionary process, except that strain HV1MA might be the only recombinant virus.

As described above, CAUSA 2.0 trees are consistent in different tree drawing mode in both of the two HIV genes, Env (Fig 5M-5P) and Gag (Fig 6E, 6F), suggesting that CAUSA 2.0 improves the phylogenetic analyses in virus. The trees inferred from PRANK and CAUSA 2.0 might be more accurate than the ClustalW trees, because they are more consistent not only in different tree drawing mode, but in different genes. Because of the underlying phylogeny-aware algorithm, PRANK is by far the best alignment method for phylogenetic analysis. CAUSA 2.0 might be equivalent to or sometimes even better than PRANK in phylogenetic analysis.

CAUSA 2.0 improves phylogenetic analyses in human and animals

The CAUSA 1.0 algorithm has been applied in POMAGO [37], a multiple genome-wide alignment tool for bacteria. To test the performance of CAUSA 2.0 in human and animals, a set of orthologous proteins families were arbitrarily selected from TreeFam-A, a manually curated database of molecular phylogenetic trees. Firstly, we aligned their CDSs with ClustalW, PRANK and CAUSA 2.0, respectively in DNA, back-translate, codon or unified mode, and then constructed phylogenetic trees using ML method, respectively in "complete deletion" (CD) mode and "using all sites" (AS) mode. Highly conserved protein families without a variable region were excluded, as we observed that different alignments and trees inferred from different methods are basically consistent with each other in those highly conserved proteins. In the variable proteins, which have at least one variable region, we observed that different MSAs and phylogenetic trees differed greatly from each other (Table S3).

As shown in Table S3, the topologies of PRANK trees are basically consistent with those of the ClustalW trees, while their bootstrap percentages are better in AS mode. Overall, in these variable protein families tested, CAUSA 2.0 trees are more consistent with the TreeFam reference trees: both the average number of consistent branches (NCBs) and bootstrap percentages (BSPs) are significantly higher than those of the other trees, including protein trees and codon trees. When all sites are used, the CAUSA 2.0-AS trees are the most consistent with the TreeFam reference trees, suggesting that the alignments aligned using CAUSA 2.0 are the most consistent between the variable region and the conserved region.

In order to avoid possible biases in human assessment, we further compared the trees with the TreeFam reference trees using TOPD/FMTS [38], a software for the evaluation of similarities between phylogenetic trees. As shown in Table S3, the TOPD/FMTS split distances (SD) are mostly consistent with the number of consistent branches counted. As shown in Table 2, CAUSA 2.0 trees shows the lowest SD in AS mode and are significantly better than those of the other alignment methods, implies that CAUSA 2.0 gives more accurate alignments, and brings a higher confidence in the downstream evolutionary analysis of the proteins and their CDSs.

Testing CAUSA 2.0 on simulated CDSs

As demonstrated in above, CAUSA 2.0 improves the alignment and evolutionary analysis of proteins in real biological data mainly due to the more accurate localization of in-frame indels which is not coincide perfectly with the codon boundaries. In nature, non-in-frame indels, or in-frame indels containing one or more stop codons, are prone to be eliminated in the struggle for existence. Therefore, they might be less frequent in real biological data than predicted. Actually, CAUSA 2.0 simulates such an evolution process by adapting indels to be in-frame as far as possible, thus makes the alignment of coding sequences biologically more significant.

We tried to quantify the accuracy of CAUSA 2.0 on simulated CDSs and compare it with those of the other alignment methods. At present, there are only a few programs, such as *indel-seq-gen* [39,40], can simulate CDSs with an adjustable indels probability (Id). Unfortunately, however, they does not distinguish in-frame indels from non-in-frame ones. As shown in Fig S3, in the CDSs simulated with low level of indels (Id = 0.01), stop codons were often generated, and all indels coincide perfectly with the codon boundaries; while in the CDSs simulated with high level of indels (Id = 0.05), a lot of non-in-frame indels and frameshift mutations were generated, makes the alignment becomes largely meaningless. In order to test the accuracy of CAUSA 2.0 and the other methods in locating in-frame and non-in-frame indels, a new DNA sequence simulator that can simulate CDSs with a controllable level of in-frame/non-in-frame indels must be developed in future studies.

Testing CAUSA 2.0 on BAliBASE

We further compared CAUSA 2.0 with the other MSA algorithm on BAliBASE, a hand-curated alignment benchmark [41-43]. As shown in Table S4, the Baliscore for the different protein alignments of BaliBase BB11001 (the high mobility group protein, HMG1) aligned by PRANK, PRANK-codon and CAUSA 2.0 are both lower than those of the other protein alignments. It has been reported that an alignment is not necessarily evolutionarily correct even if it is structurally accurate[7]. Therefore, BAliBASE, which is based on structure alignments, might not be suitable for the assessment of alignment methods aimed at creating evolutionarily correct alignments, such as PRANK, codon alignment or CAUSA 2.0.

The information contents of different alignments

The information content (IC) for DNA and protein sequences were given by Shannon entropy [44-46]. For unaligned DNA or protein sequences, IC is computed by the frequencies of different states occurred in the sequences (4 bases, 20 amino acids or 64 codons).

$$S_B =- \sum_{i=1}^{4} P_i \, log_{\,2}(P_i) \qquad [1]$$

$$S_A =- \sum_{j=1}^{20} P_j \, log_{\,2}(P_j) \qquad [2]$$

Given a triplet codon, let P_k, P_l, P_m be the frequencies of the three bases,

$$S_C =- \sum_{k=1}^{4} \sum_{l=1}^{4} \sum_{m=1}^{4} P_k P_l P_m \, log_{\,2}(P_k P_l P_m) \qquad [3]$$

The information contents of DNA and protein sequences reach their upper limits when different states occur in equal frequencies:

$$\mathrm{Max}(\mathrm{S_A}) = \log_2\left(\frac{1}{4}\right) = 2.0000$$

$$\mathrm{Max}(\mathrm{S_B}) = \log_2\left(\frac{1}{20}\right) = 4.3219$$

$$\mathrm{Max}(\mathrm{S_C}) = \log_2\left(\frac{1}{64}\right) = 6.0000$$

After the sequences were aligned, the entropy of the aligned sequences will change, as an additional state (the gap-state) is introduced. Let S_G be the increased IC comes from the gaps inserted during the alignment process, we have:

$$S_G =- P_g \log_2(P_g) \qquad [4]$$

$$S_{Base} = S_B + S_G \qquad [5]$$

$$S_{AA} = S_A + S_G \qquad [6]$$

$$S_{Codon} = S_C + S_G \qquad [7]$$

Where P_i, P_j, P_k, P_g is the frequencies of amino acids (A), bases (B), codons (C) and gaps (G) in the aligned sequences, respectively. Given a set of random sequences, in theory the IC of protein sequences is higher than that of DNA sequences, and that of coding sequences is even higher. Therefore, a protein alignment would be better than the DNA alignment, and the corresponding codon alignment could be even better.

In DNA, protein, codon or back translated protein alignments, there is only one gap-state because codons are indivisible; however, there are many more gap-states in CAUSA 2.0, because every base of a codon can be replaced by a gap. Given a triplet codon, let P_k, P_l, P_m be the frequencies of the three bases, and P_g be the frequency of a base to be replaced by a gap. Then,

$$P_0 = P_k P_l P_m (1 - P_g)^3 \qquad [8]$$

$$P_1 = P_k P_l P_m P_g (1 - P_g)^2 \qquad [9]$$

$$P_2 = P_k P_l P_m P_g^2 (1 - P_g) \qquad [10]$$

$$P_3 = P_k P_l P_m P_g^3 \qquad [11]$$

Where P_1, P_2, P_3 is the probability of a codon contains 1, 2 and 3 gaps, respectively. The total probability of split codons (the codons contain at least one gaps) is,

$$P_s = P_1 + P_2 + P_3 = P_k P_l P_m P_g (1 - P_g + P_g^2) \qquad [12]$$

The information content of a unified alignment is given by,

$$S_u = S_C - \sum_{k=1}^{4} \sum_{l=1}^{4} \sum_{m=1}^{4} P_s \, log_{\,2}(P_s) \qquad [13]$$

In an unified alignment: (1) If $P_g = 0$, then $P_s = 0$, $S_u = S_{Codon}$, there is no split codon, and thus the information content of the unified alignment equals to that of a codon alignment; In other words, codon alignment could be considered as a special case of the unified alignment, where no split codon is present. (2) If $P_g > 0$, then $P_s > 0$, $S_u > S_{Codon}$, the split codons enhance the information content of the unified alignment, and makes it greater than those of the DNA, protein or codon alignments, thus explains why a unified alignment could be more accurate than the other level of alignments.

The computational efficiency of CAUSA 2.0

Among all the programs tested, as shown in Table S5, MAFFT is the fastest, especially when the number of sequences and the total length increases; MUSCLE, an iterative method, is faster than ClustalW. T-Coffee is even faster as it adopted a multiple-threading paralleled computation. PRANK, however, is very time-consuming: even in moderate-sized datasets (20 ~ 50KB in total length), while other software finished aligning in a few minutes, it takes PRANK several hours to align them. PRANK codon alignment is even more time-consuming, probably due to its adopting of a large (61x61) scoring matrix. In contrast, using a small (24x24) scoring matrix, CAUSA 2.0 is computationally very efficient. The computation time of CAUSA 2.0 is similar to that of ClustalW. However, since CAUSA 2.0 outputs both protein and codon alignment in a single run, it is considered more efficient than most of the other programs.

Discussion and conclusion

Multiple sequence alignment is widely useful in bioinformatics, molecular evolution, genetics and genomics, such as reconstruction of phylogenetic history, structure and functional analyses. Moreover, phylogenetic trees inferred from DNA, protein and codon alignments are often inconsistent. It has been believed that the phylogenetic signal disappears more rapidly from DNA sequences than from their encoded proteins, and therefore it has been preferable to align protein and their CDSs at AA level [8]. However, some important information carried by CDSs, including synonymous substitutions, codon splitting/fusion, siDel and ipFS, get lost when they were translated into AAs, and makes the alignments and phylogenetic trees sometimes inaccurate. CAUSA 2.0 performs a two level dynamic programming alignment algorithm at codon and AA level. CAUSA 2.0 minimizes the total amount of mutations at both codon and AA level efficiently, locates the in-frame indels precisely and better interprets their role in the molecular evolution. In CAUSA 2.0, the boundary of every triplet codon is defined by their encoded amino acid without forcing them into indivisible units. Thanks to the position constraint inherent in the codon-aa 4-tuples, CAUSA 2.0 aligns codons while keeps the corresponding AAs staying in frame without using post-alignment adjustments.

In the previous study [27], we have evaluated CAUSA 1.0 for the phylogenetic analysis in virus, bacteria and mammalian animals. However, we used the combined alignments directly for the construction of the phylogenetic trees, and the trees were inferred based on p-distances as input to neighbor joining (NJ) method. It has been pointed out that there are several defects in that study, mainly including: (1) Although CAUSA does improve the alignment of CDSs, there is no model to support using combined alignment directly for the construction of phylogenetic trees; (2) The phylogenetic trees were constructed using Neighbor Joining (NJ) method, but the use of p-distances for phylogenetic analysis has been shown to be systematically biased, because the p-distances are known to undercount the number of substitutions between a pair of sequences, and perform particularly poorly for distantly related sequences; (3) CAUSA is compared only with protein and DNA level alignments, but not with codon level alignments.

CAUSA 2.0 is an major update of the CAUSA algorithm, and the present study differs from the previous one mainly in the following aspects: (1) Instead of using a combined alignment directly for the construction of phylogenetic trees, the combined alignment was separated into a protein alignment and a codon alignment, and used to construct the phylogenetic trees, therefore any existing model can still apply in the subsequent evolutionary analysis; (2) The phylogenetic trees were inferred using a Maximum Likelihood (ML) method in the MEGA software; (3) CAUSA 2.0 was compared not only with protein and DNA alignments, but also with codon alignments aligned by PRANK. (4) A mathematical model based on entropy information content was developed, which explains why CAUSA 2.0 works better than DNA, protein and codon level alignments; (5) Several newly discovered mutation mechanisms linked to certain kind of genetic diseases were discussed; (6) It was showed that the computational efficiency of the approach is higher than the other methods.

Based on the above analysis, it is concluded that CAUSA 2.0 gives more accurate alignment for proteins and their CDSs, as well as more confident predictions in the phylogenetic and evolutionary analysis. The main benefit of this method is the ability to detect in-frame, as well as non in-frame indels, while keeping the respective amino acids aligned. CAUSA 2.0 is also superior in computational efficiency when compared with the other approaches.

Finally, CAUSA 2.0 programs run in Microsoft Windows and Linux, written respectively in Microsoft Visual C# and Java, are available for free in the website www.DNAPlusPro.com.

Acknowledgements

We would like to thank Shu Cai for his contribution in the programming of the present version of this software. We also would like to thank Yu Fu, Qi Wang, Chandra Sekhar Pedamallu, from the Broad Institute and Dana-Farber Cancer Institute (DFCI), and Shuang-yong Xu from New England Biolabs, Inc. for their contribution in the development of the earlier version of this software.

References

1 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673-4680 (1994).

2 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

3 Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. *Methods in molecular biology* **1079**, 131-146, doi:10.1007/978-1-62703-646-7_8 (2014).

4 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

5 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* **302**, 205-217, doi:10.1006/jmbi.2000.4042 (2000).

6 Loytynoja, A. Phylogeny-aware alignment with PRANK. *Methods in molecular biology* **1079**, 155-170, doi:10.1007/978-1-62703-646-7_10 (2014).

7 Loytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632-1635, doi:10.1126/science.1158395 (2008).

8 Wernersson, R. & Pedersen, A. G. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic acids research* **31**, 3537-3539 (2003).

9 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-10919 (1992).

10 Styczynski, M. P., Jensen, K. L., Rigoutsos, I. & Stephanopoulos, G. BLOSUM62 miscalculations improve search performance. *Nature biotechnology* **26**, 274-275, doi:10.1038/nbt0308-274 (2008).

11 Gonnet, G. H., Cohen, M. A. & Benner, S. A. Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443-1445 (1992).

12 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34**, W609-612, doi:10.1093/nar/gkl315 (2006).

13 Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic acids research* **38**, W7-13, doi:10.1093/nar/gkq291 (2010).

14 Hein, J. An algorithm combining DNA and protein alignment. *Journal of theoretical biology* **167**, 169-174, doi:10.1006/jtbi.1994.1062 (1994).

15 Hein, J. & Stovlbaek, J. Combined DNA and protein alignment. *Methods in enzymology* **266**, 402-418 (1996).

16 Schneider, A., Cannarozzi, G. M. & Gonnet, G. H. Empirical codon substitution matrix. *BMC bioinformatics* **6**, 134, doi:10.1186/1471-2105-6-134 (2005).

17 Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution* **11**, 725-736 (1994).

18 McClellan, D. A. The codon-degeneracy model of molecular evolution. *Journal of molecular evolution* **50**, 131-140 (2000).

19 Zaheri, M., Dib, L. & Salamin, N. A generalized mechanistic codon model. *Molecular biology and evolution* **31**, 2528-2541, doi:10.1093/molbev/msu196 (2014).

20 Miyazawa, S. Superiority of a mechanistic codon substitution model even for protein sequences in phylogenetic analysis. *BMC evolutionary biology* **13**, 257, doi:10.1186/1471-2148-13-257 (2013).

21 McClellan, D. A. The phylogenetic utility of the codon-degeneracy model. *Journal of molecular evolution* **51**, 185-193 (2000).

22 Abby, S. S., Tannier, E., Gouy, M. & Daubin, V. Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 4962-4967, doi:10.1073/pnas.1116871109 (2012).

23 Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10557-10562, doi:10.1073/pnas.0409137102 (2005).

24 Szalkowski, A. M. Fast and robust multiple sequence alignment with phylogeny-aware gap placement. *BMC bioinformatics* **13**, 129, doi:10.1186/1471-2105-13-129 (2012).

25 Loytynoja, A., Vilella, A. J. & Goldman, N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* **28**, 1684-1691, doi:10.1093/bioinformatics/bts198 (2012).

26 Loytynoja, A. & Goldman, N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC bioinformatics* **11**, 579, doi:10.1186/1471-2105-11-579 (2010).

27 Wang, X. *et al.* Accurate reconstruction of molecular phylogenies for proteins using codon and amino acid unified sequence alignments (CAUSA). *Nature Precedings* <*http://hdl.handle.net/10101/npre.2011.6730.1*> (2011).

28 Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Human molecular genetics* **19**, R131-136, doi:10.1093/hmg/ddq400 (2010).

29 Amor, I. M. *et al.* Severe osteogenesis imperfecta caused by a small in-frame deletion in CRTAP. *American journal of medical genetics. Part A* **155A**, 2865-2870, doi:10.1002/ajmg.a.34269 (2011).

30 Guillaud Bataille, M. *et al.* Systematic screening for PRKAR1A gene rearrangement in Carney complex: identification and functional characterization of a new in-frame deletion. *European journal of endocrinology / European Federation of Endocrine Societies* **170**, 151-160, doi:10.1530/EJE-13-0740 (2014).

31 Yang, W., Wei, H. & Sang, Y. KCNJ11 in-frame 15-bp deletion leading to glibenclamide-responsive neonatal diabetes mellitus in a Chinese child. *Journal of pediatric endocrinology & metabolism : JPEM* **26**, 743-746 (2013).

32 Yang, S. Y. *et al.* EGFR exon 19 in-frame deletion and polymorphisms of DNA repair genes in never-smoking female lung adenocarcinoma patients. *International journal of cancer. Journal international du cancer* **132**, 449-458, doi:10.1002/ijc.27630 (2013).

33 Weedon, M. N. *et al.* An in-frame deletion at the polymerase active site of POLD1 causes a multisystem disorder with lipodystrophy. *Nature genetics* **45**, 947-950, doi:10.1038/ng.2670 (2013).

34 Dudkiewicz, M., Szczepinska, T., Grynberg, M. & Pawlowski, K. A novel protein kinase-like domain in a selenoprotein, widespread in the tree of life. *PloS one* **7**, e32138, doi:10.1371/journal.pone.0032138 (2012).

35 Sobhani, M., Tabatabaiefar, M. A., Rajab, A., Kajbafzadeh, A. M. & Noori-Daloii, M. R. Molecular characterization of WFS1 in an Iranian family with Wolfram syndrome reveals a novel frameshift mutation associated with early symptoms. *Gene* **528**, 309-313, doi:10.1016/j.gene.2013.06.040 (2013).

36 Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327-331, doi:10.1038/nature12130 (2013).

37 Wieseke, N., Lechner, M., Ludwig, M. & Marz, M. in *Bioinformatics Research and Applications* 249-260 (Springer, 2013).

38 Puigbo, P., Garcia-Vallve, S. & McInerney, J. O. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics* **23**, 1556-1558, doi:10.1093/bioinformatics/btm135 (2007).

39 Strope, C. L., Abel, K., Scott, S. D. & Moriyama, E. N. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Molecular biology and evolution* **26**, 2581-2593, doi:10.1093/molbev/msp174 (2009).

40 Strope, C. L., Scott, S. D. & Moriyama, E. N. indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels. *Molecular biology and evolution* **24**, 640-649, doi:10.1093/molbev/msl195 (2007).

41 Thompson, J. D., Koehl, P., Ripp, R. & Poch, O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* **61**, 127-136, doi:10.1002/prot.20527 (2005).

42 Bahr, A., Thompson, J. D., Thierry, J. C. & Poch, O. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic acids research* **29**, 323-326 (2001).

43 Thompson, J. D., Plewniak, F. & Poch, O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**, 87-88 (1999).

44 Rackovsky, S. & Scheraga, H. A. On the information content of protein sequences. *Journal of biomolecular structure & dynamics* **28**, 593-594; discussion 669-674, doi:10.1080/073911011010524957 (2011).

45 Weiss, O., Jimenez-Montano, M. A. & Herzel, H. Information content of protein sequences. *Journal of theoretical biology* **206**, 379-386, doi:10.1006/jtbi.2000.2138 (2000).

46 Wu, T. T., Fitch, W. M. & Margoliash, E. The information content of protein amino acid sequences. *Annual review of biochemistry* **43**, 539-566, doi:10.1146/annurev.bi.43.070174.002543 (1974).

Table 1(on next page)

Some disease-related indels with codon splitting/fusion events derived from Human Gene Mutation Database.

Accession Number	Deletion (^codon number)	Disease	Reference
CD031549	GCC GTG^^{177}CGC Aag gca gCC CTG GTC AT	WS	Colosimo (2003) Hum Mutat 21,622
CD011664	CTG GTC^^{349}ATC Ttc tAC CTG TCC TT	WS	Khanim (2001) Hum Mutat 17,357
CD058021	TTC GTC^^{416}ATC Ttc tCC TTC CCC AT	WS	Fang (2005) Zhonghua Yi Xue Za Zhi 85,2468
CD050485	CC GCC^^{466}GGC CTg ctA TCG CTG CTG	WS	Giuliano (2005) Hum Mutat 25,99
CD021029	CCC TGC^^{766}CAC Atc aAG AAG TTC GA	WS	Cryns (2002) Hum Genet 110,389
CD983013	GC GTC^^{504}CCG TGc ctg ctc tat gtc taC CTG CTC TAT	WS	Inoue (1998) Nat Genet 20,143
CD983491	GG GCC^^{459}CTG GCc acc gag gtC ACC GCC GGC	WS	Strom (1998) Hum Mol Genet 7,2021
CD031551	GC TTC^^{539}ATG TGg tgT GAG CTC TCC	WS	Colosimo (2003) Hum Mutat 21,622
CD962054	GTG GTG^^{166}TTC Ttcg GGA CGG AGT A	LQTS	Wang (1996) Nat Genet 12,17
CD076814	GGC CTC^^{274}ATC Ttc tCC TCG TAC TT	LQTS	Aizawa (2007) J CaRdiovasc ElectrOphysiol 18,972
CD044136	TC ATC^^{275}TTC TCc tcG TAC TTT GTG	LQTS	Gouas (2004) Cardiovasc Res 63,60
CD097252	AC CGA^^{595}GTA GAa gaC AAG GTA GGC	LQTS	Kapplinger (2009) Heart Rhythm 6,1297
CD097261	AC ATG^^{613}CTT CAc caG CTG CTC TCC	LQTS	Kapplinger (2009) Heart Rhythm 6,1297
CD033996	GC ATT^^{382}AAC CCa att gct ctG TAT TTG GTG	HD	Garcia-Barcelo (2003) Clin Chem 50,93
CI031592	GAG GAC^^{797}GAC Gga cTC ACC AAG GA	WS	Colosimo (2003) Hum Mutat 21,622
CI087227	AC GCG^^{53}CCC ATc gcg ccc atC GCG CCC GGC	LQTS	Berge (2008) Scand J Clin Lab Invest 68,362
CI013349	GG CAG^^{362}AAG Cag aag caC TTC AAC CGG	LQTS	Kapplinger (2009) Heart Rhythm 6,1297
CI983230	GCC ACG^^{1509}GCT Tcg gct tCC ATT GAC AT	HI	Nestorowicz (1998) Hum Mol Genet 7,1119

Note: WS - Wolfram Syndrome; LQTS - Long QT syndrome; HD- Hirschsprung disease; HI – Hyperinsulinism.

Table 2(on next page)

Comparing the phylogenetic trees of 30 protein families with their TreeFam reference trees by human eyes and using TOPD/FMTS.

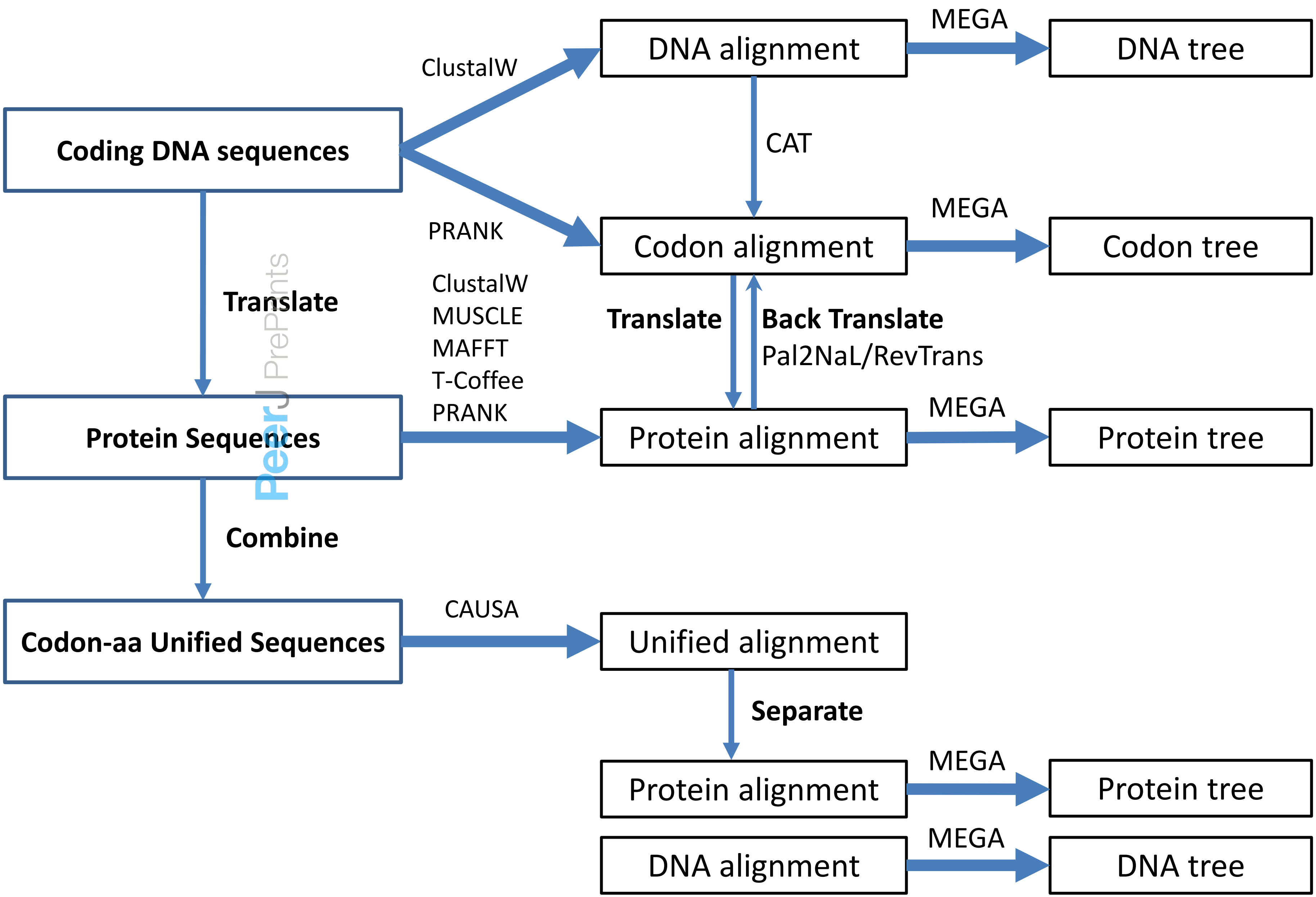
PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1214v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

Software	Alignment Method	Gap or Missing data treatment	Average NCB (Human)	Average SD (TOPD/FMTS)	T-Test Significance
ClustalW	ClustalW-DNA	CD	11.37	0.1777	**
		AS	11.67	0.1498	*
	ClustalW-codon (Back translate)	CD	11.67	0.1512	*
		AS	11.83	0.1332	*
PRANK	PRANK-DNA	CD	11.33	0.1869	**
		AS	11.37	0.1719	**
	PRANK-codon	CD	10.47	0.3031	**
		AS	11.40	0.1865	**
CAUSA 2.0	CAUSA 2.0-DNA	CD	11.47	0.1832	**
		AS	**11.97**	**0.1156**	

Note: NCB - Number of Consistent Branches; SD –TOPD/FMTS Split Distance; CD - Complete deletion; AS - Use all sites; "*" - Significant difference (P<0.05); "**" –Extreme significant difference (P<0.01).

Figure 1(on next page)

The working flowchart of different strategies for aligning proteins and their coding DNA sequences.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1214v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

Figure 2(on next page)

The protein views of different alignments of a variable region (V2) of Env. HIV or SIV strains were derived from the seed alignment of Pfam gp120 protein family (pf00516).

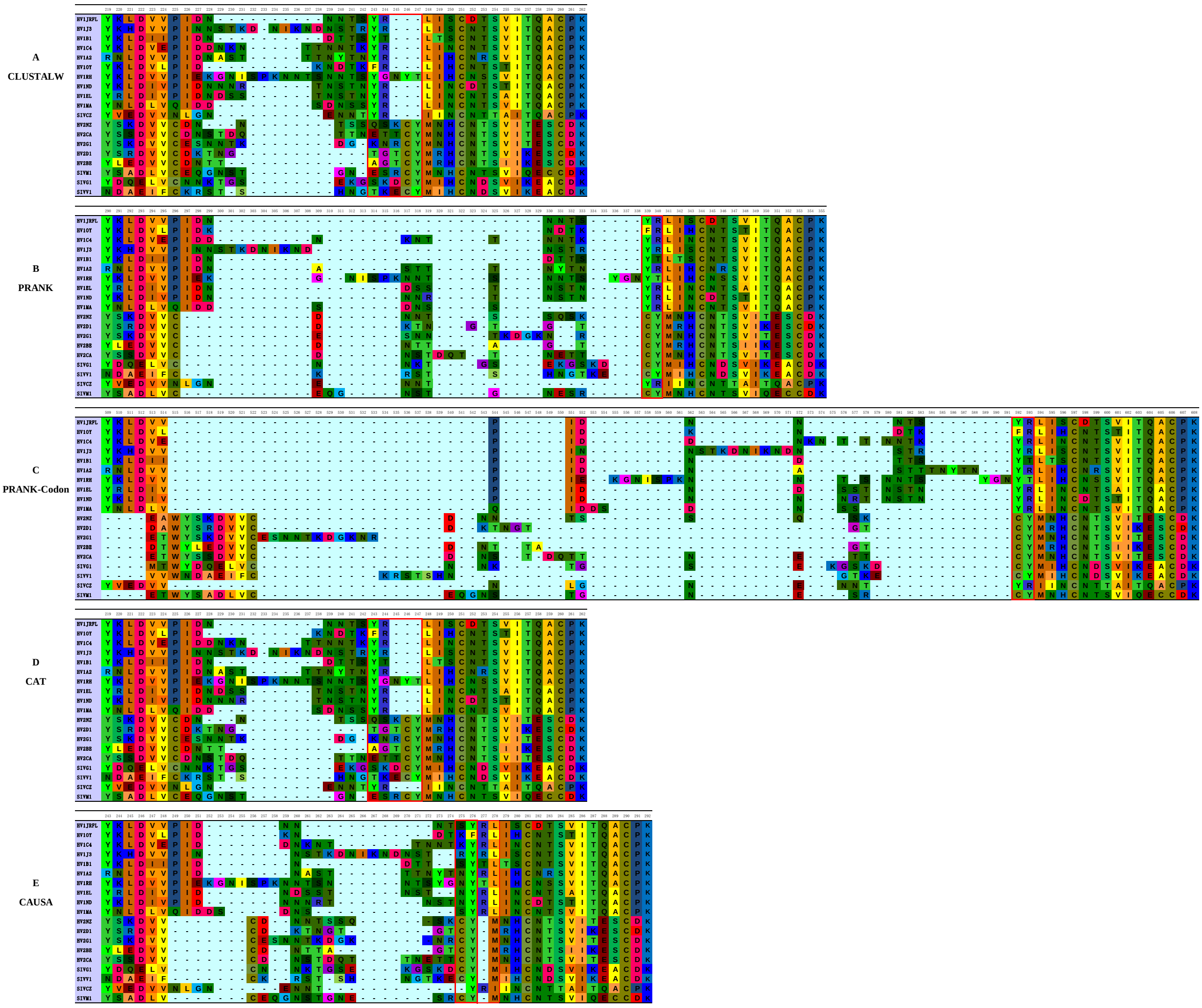

A
CLUSTALW
B
PRANK
C
PRANK-Codon
D
CAT
E
CAUSA

Figure 3(on next page)

Localization of in-frame indels in different alignment programs. (A, B, C) Codon splitting; (D, E, G) Codon fusion; (G, H, I) siDel and ipFS.

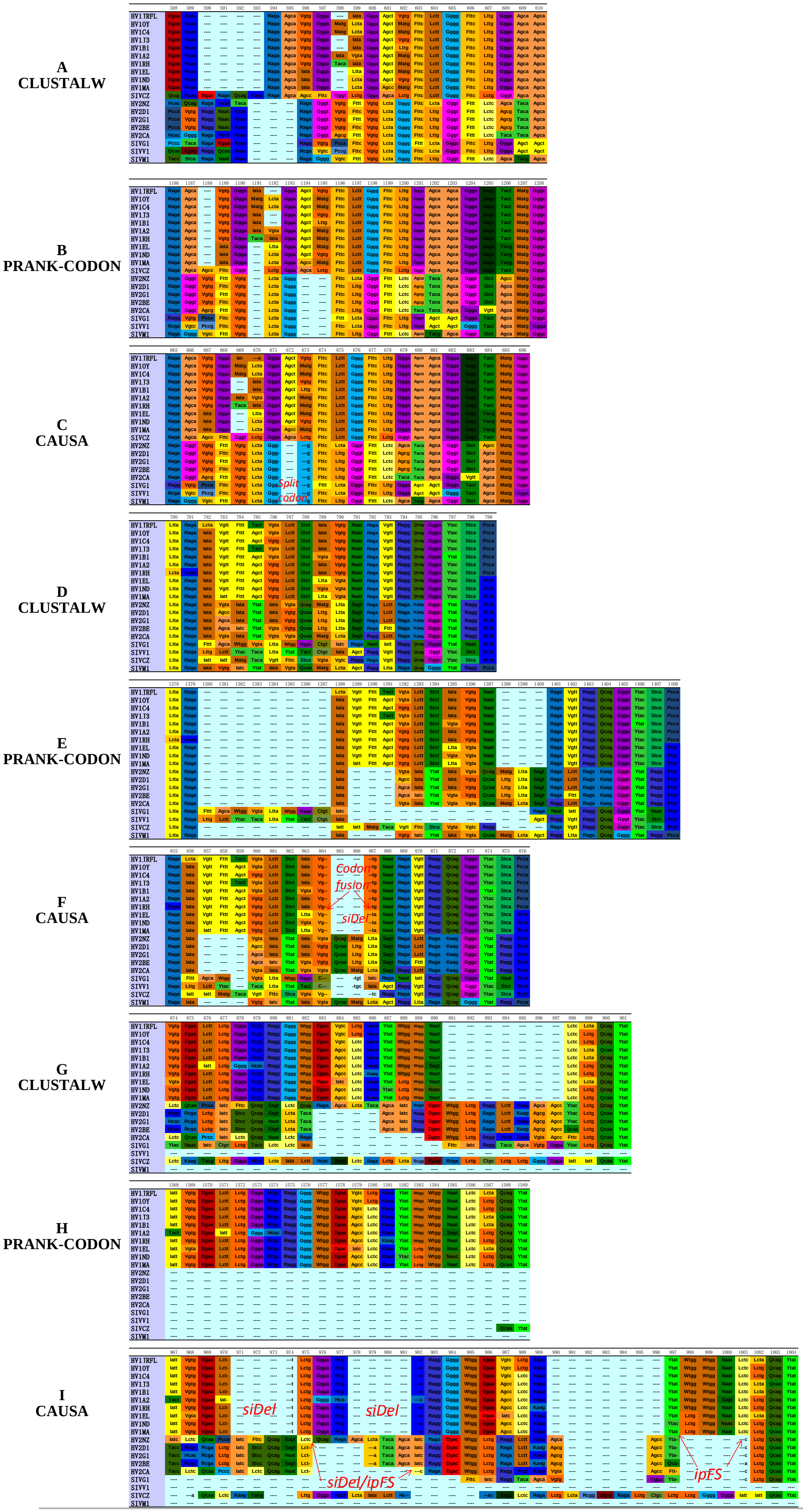

A
CLUSTALW
B
PRANK-CODON
C
CAUSA
Split
codon
D
CLUSTALW
E
PRANK-CODON
F
CAUSA
Codon
fusion
siDel
G
CLUSTALW
H
PRANK-CODON
I
CAUSA
siDel
siDel
siDel/ipFS
ipFS
HV1TRFL
HV1OY
HV1C4
HV1I3
HV1B1
HV1A2
HV1RH
HV1EL
HV1ND
HV1MA
SIVCZ
HV2NZ
HV2D1
HV2G1
HV2BE
HV2CA
SIVG1
SIVV1
SIVM1

Figure 4(on next page)

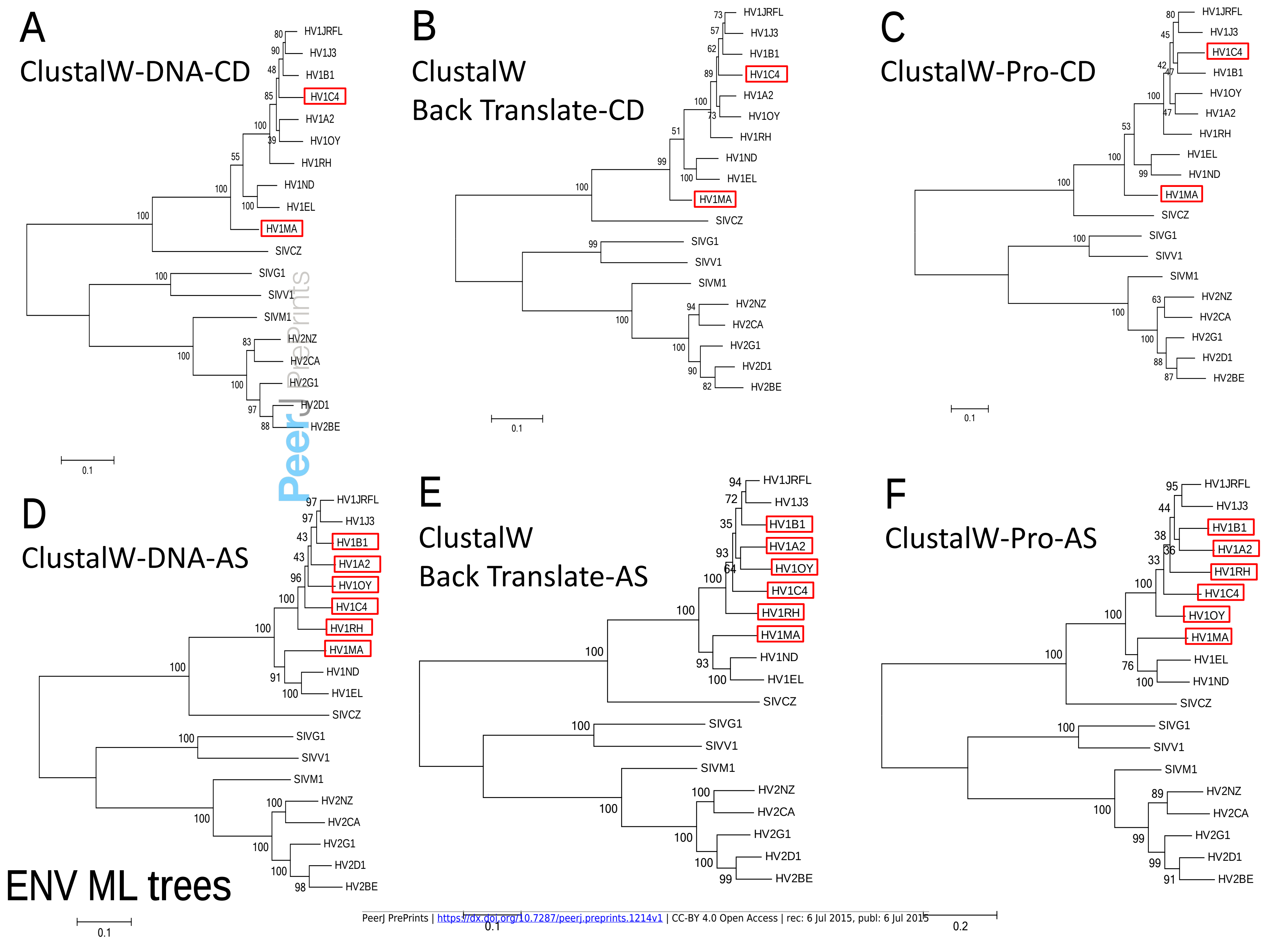
The CAUSA 2.0 view of an in-frame deletion related to Wolfram Syndrome (CD031549), showing a codon fusion event.

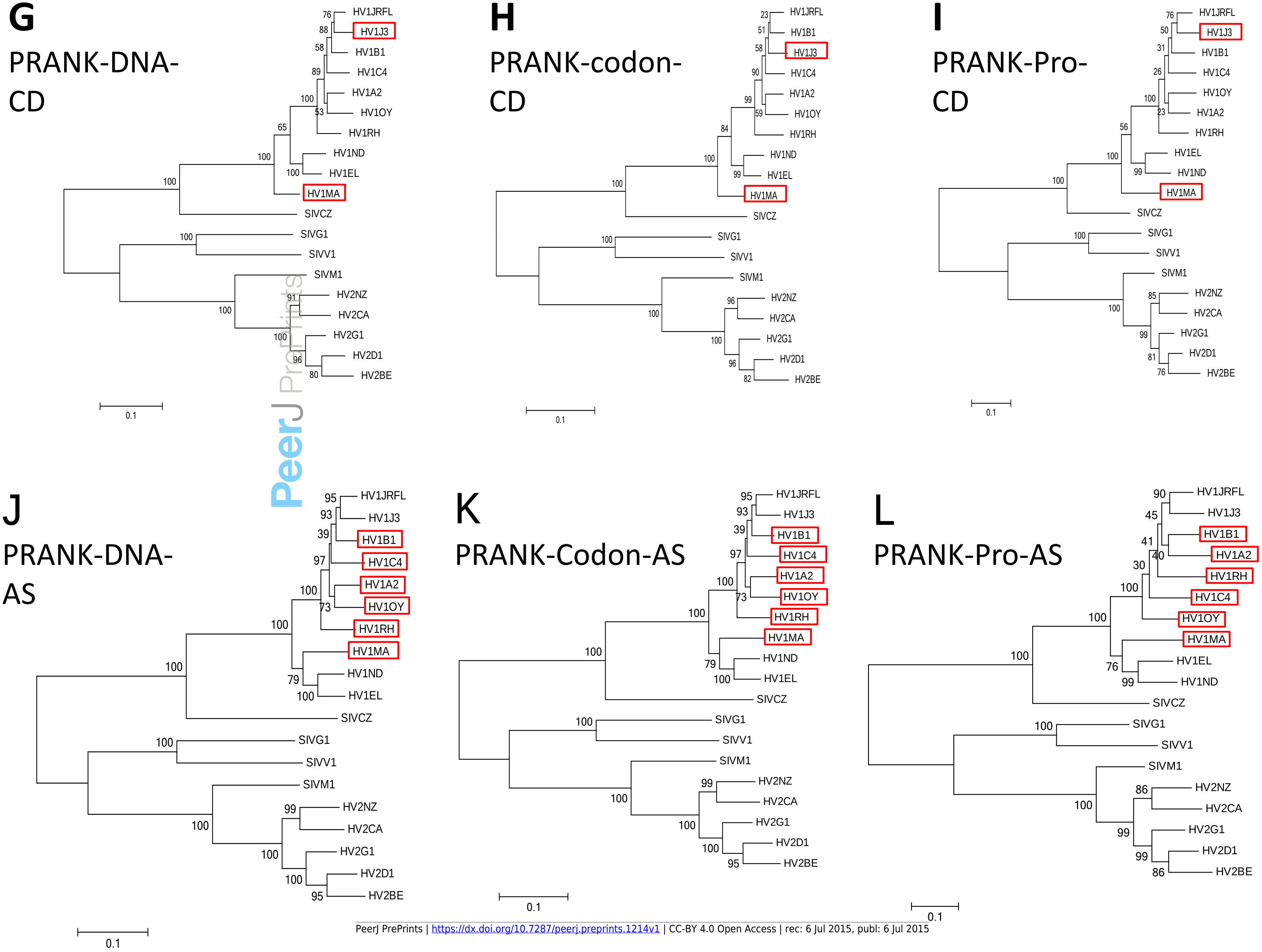
CD031549 (deletion)

Seq Name	1	2	3	4	5	6	7	8
Normal	Agcc	Vgtg	Rcgc	Kaag	Agca	Agcc	Lctg	Vgtc
Wolfram 1	Agcc	Vgtg	Rcgc	Ta--	----	--cc	Lctg	Vgtc
Wolfram 2	Agcc	Vgtg	Rcgc	Ta--	----	--cc	Lctg	Vgtc
Normal	A	V	R	K	A	A	L	V
Wolfram 1	A	V	R	T	-	-	L	V
Wolfram 2	A	V	R	T	-	-	L	V
Normal	gcc	gtg	cgc	aag	gca	gcc	ctg	gtc
Wolfram 1	gcc	gtg	cgc	a--	---	-cc	ctg	gtc
Wolfram 2	gcc	gtg	cgc	a--	---	-cc	ctg	gtc

Figure 5(on next page)

The maximum likelihood trees for HIV Env protein inferred from different alignments and by different gap/missing data treatment (CD-complete deletion; AS-Use all sites).


A
ClustalW-DNA-CD
80 HV1JRFL
90 HV1J3
48 HV1B1
85 HV1C4
100 HV1A2
39 HV1OY
HV1RH
55
100 HV1ND
100 HV1EL
HV1MA
100
SIVCZ
100 SIVG1
SIVV1
SIVM1
83 HV2NZ
HV2CA
100
100 HV2G1
97 HV2D1
88 HV2BE
0.1
B
ClustalW
Back Translate-CD
73 HV1JRFL
57 HV1J3
62 HV1B1
89 HV1C4
100 HV1A2
73 HV1OY
HV1RH
51
99 HV1ND
100 HV1EL
HV1MA
100
SIVCZ
99 SIVG1
SIVV1
SIVM1
94 HV2NZ
HV2CA
100
100 HV2G1
90 HV2D1
82 HV2BE
0.1
C
ClustalW-Pro-CD
80 HV1JRFL
45 HV1J3
HV1C4
42
47 HV1B1
100 HV1OY
47 HV1A2
HV1RH
53
100 HV1EL
99 HV1ND
HV1MA
100
SIVCZ
100 SIVG1
SIVV1
SIVM1
63 HV2NZ
HV2CA
100
100 HV2G1
88 HV2D1
87 HV2BE
0.1
D
ClustalW-DNA-AS
97 HV1JRFL
97 HV1J3
43 HV1B1
43 HV1A2
96 HV1OY
100 HV1C4
100 HV1RH
HV1MA
100
91 HV1ND
100 HV1EL
SIVCZ
100 SIVG1
SIVV1
SIVM1
100 HV2NZ
HV2CA
100
100 HV2G1
100 HV2D1
98 HV2BE
ENV ML trees
0.1
E
ClustalW
Back Translate-AS
94 HV1JRFL
72 HV1J3
35 HV1B1
93 HV1A2
64 HV1OY
100 HV1C4
100 HV1RH
HV1MA
100
93 HV1ND
100 HV1EL
SIVCZ
100 SIVG1
SIVV1
SIVM1
100 HV2NZ
HV2CA
100
100 HV2G1
100 HV2D1
99 HV2BE
0.1
F
ClustalW-Pro-AS
95 HV1JRFL
44 HV1J3
38 HV1B1
36 HV1A2
33 HV1RH
100 HV1C4
100 HV1OY
HV1MA
100
76 HV1EL
100 HV1ND
SIVCZ
100 SIVG1
SIVV1
SIVM1
89 HV2NZ
HV2CA
100
99 HV2G1
99 HV2D1
91 HV2BE
0.2


G
PRANK-DNA-CD
HV1JRFL
HV1J3
HV1B1
HV1C4
HV1A2
HV1OY
HV1RH
HV1ND
HV1EL
HV1MA
SIVCZ
SIVG1
SIVV1
SIVM1
HV2NZ
HV2CA
HV2G1
HV2D1
HV2BE
76
88
58
89
100
53
65
100
100
100
100
100
91
100
100
96
80
0.1
H
PRANK-codon-CD
HV1JRFL
HV1B1
HV1J3
HV1C4
HV1A2
HV1OY
HV1RH
HV1ND
HV1EL
HV1MA
SIVCZ
SIVG1
SIVV1
SIVM1
HV2NZ
HV2CA
HV2G1
HV2D1
HV2BE
23
51
58
90
99
59
84
100
99
100
100
100
96
100
96
82
0.1
I
PRANK-Pro-CD
HV1JRFL
HV1J3
HV1B1
HV1C4
HV1OY
HV1A2
HV1RH
HV1EL
HV1ND
HV1MA
SIVCZ
SIVG1
SIVV1
SIVM1
HV2NZ
HV2CA
HV2G1
HV2D1
HV2BE
76
50
31
26
100
23
56
100
99
100
100
85
100
99
81
76
0.1
J
PRANK-DNA-AS
HV1JRFL
HV1J3
HV1B1
HV1C4
HV1A2
HV1OY
HV1RH
HV1MA
HV1ND
HV1EL
SIVCZ
SIVG1
SIVV1
SIVM1
HV2NZ
HV2CA
HV2G1
HV2D1
HV2BE
95
93
39
97
100
73
100
79
100
100
100
99
100
100
100
95
0.1
K
PRANK-Codon-AS
HV1JRFL
HV1J3
HV1B1
HV1C4
HV1A2
HV1OY
HV1RH
HV1MA
HV1ND
HV1EL
SIVCZ
SIVG1
SIVV1
SIVM1
HV2NZ
HV2CA
HV2G1
HV2D1
HV2BE
95
93
39
97
100
73
100
79
100
100
100
99
100
100
100
95
0.1
L
PRANK-Pro-AS
HV1JRFL
HV1J3
HV1B1
HV1A2
HV1RH
HV1C4
HV1OY
HV1MA
HV1EL
HV1ND
SIVCZ
SIVG1
SIVV1
SIVM1
HV2NZ
HV2CA
HV2G1
HV2D1
HV2BE
90
45
41
40
30
100
100
76
99
100
100
86
100
99
99
86
0.1

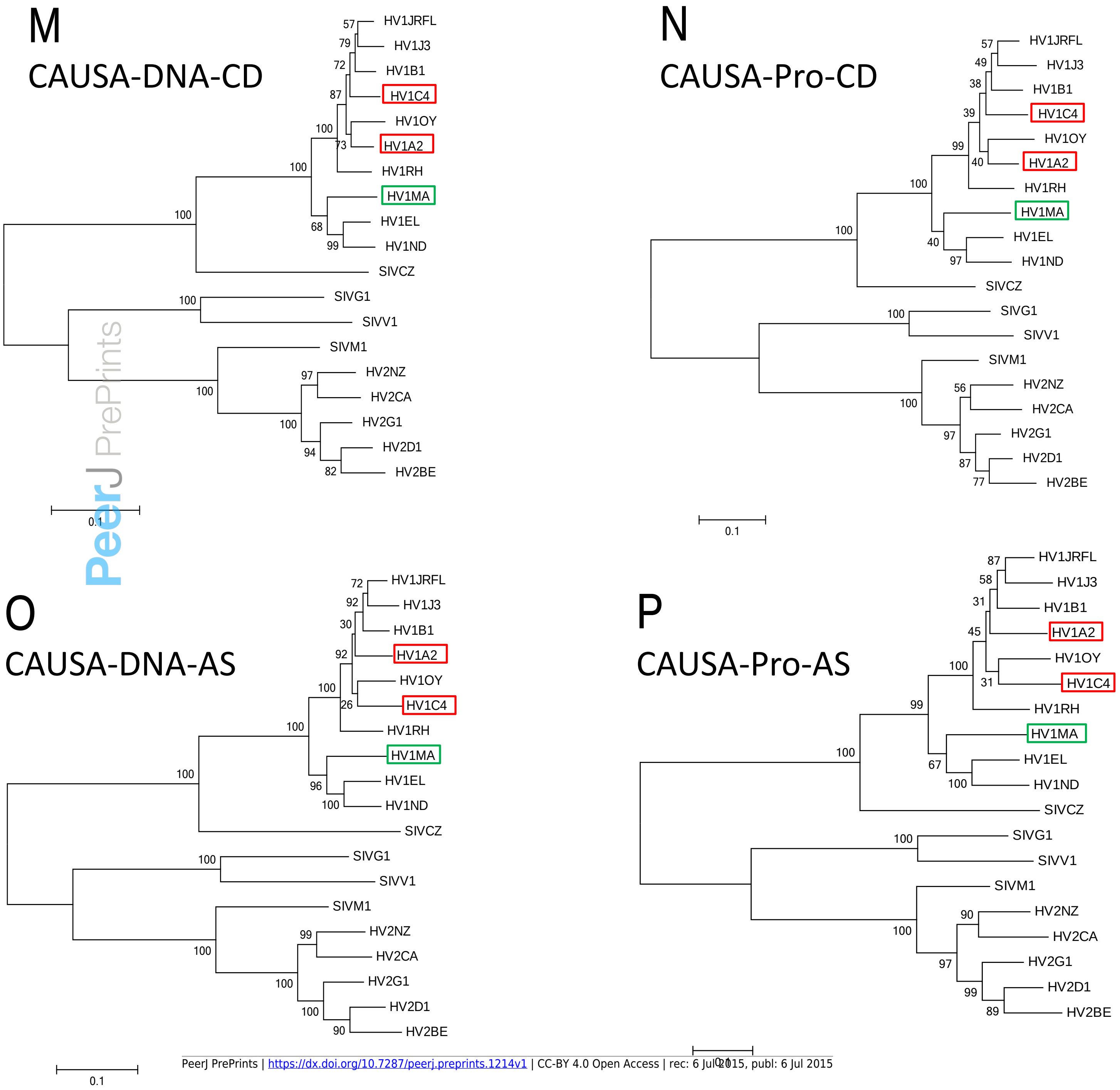
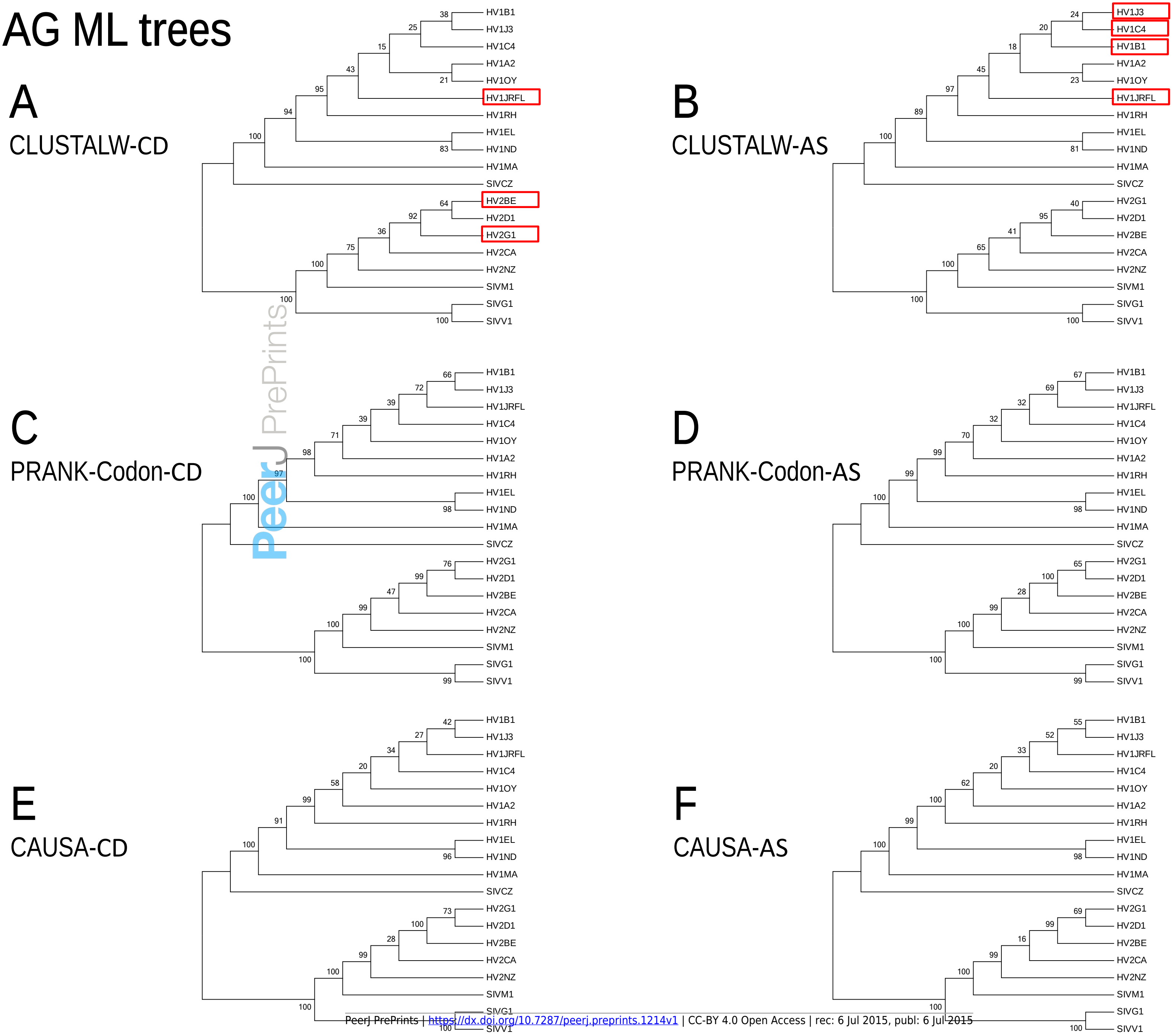

M
CAUSA-DNA-CD
57
HV1JRFL
79
HV1J3
72
HV1B1
87
HV1C4
100
HV1OY
73
HV1A2
100
HV1RH
HV1MA
100
68
HV1EL
99
HV1ND
SIVCZ
100
SIVG1
SIVV1
SIVM1
97
HV2NZ
100
HV2CA
100
HV2G1
HV2D1
94
82
HV2BE
0.1
N
CAUSA-Pro-CD
57
HV1JRFL
49
HV1J3
38
HV1B1
39
HV1C4
99
HV1OY
40
HV1A2
100
HV1RH
HV1MA
100
40
HV1EL
97
HV1ND
SIVCZ
100
SIVG1
SIVV1
SIVM1
56
HV2NZ
100
HV2CA
97
HV2G1
87
HV2D1
77
HV2BE
0.1
O
CAUSA-DNA-AS
72
HV1JRFL
92
HV1J3
30
HV1B1
92
HV1A2
100
HV1OY
26
HV1C4
100
HV1RH
HV1MA
100
96
HV1EL
100
HV1ND
SIVCZ
100
SIVG1
SIVV1
SIVM1
99
HV2NZ
100
HV2CA
100
HV2G1
100
HV2D1
90
HV2BE
0.1
P
CAUSA-Pro-AS
87
HV1JRFL
58
HV1J3
31
HV1B1
45
HV1A2
100
HV1OY
31
HV1C4
99
HV1RH
HV1MA
100
67
HV1EL
100
HV1ND
SIVCZ
100
SIVG1
SIVV1
SIVM1
90
HV2NZ
100
HV2CA
97
HV2G1
99
HV2D1
89
HV2BE
0.1

Figure 6(on next page)

The maximum likelihood trees for HIV GAG protein inferred from different alignments and by different gap/missing data treatment.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1214v1 | CC-BY 4.0 Open Access | rec: 6 Jul 2015, publ: 6 Jul 2015

GAG ML trees


A
CLUSTALW-CD
38
HV1B1
HV1J3
25
HV1C4
15
HV1A2
43
21
HV1OY
HV1JRFL
95
HV1RH
94
HV1EL
100
83
HV1ND
HV1MA
SIVCZ
64
HV2BE
HV2D1
92
36
HV2G1
75
HV2CA
100
HV2NZ
SIVM1
100
SIVG1
100
SIVV1
B
CLUSTALW-AS
24
HV1J3
HV1C4
20
HV1B1
18
HV1A2
45
23
HV1OY
97
HV1JRFL
89
HV1RH
HV1EL
100
81
HV1ND
HV1MA
SIVCZ
40
HV2G1
HV2D1
95
HV2BE
41
65
HV2CA
100
HV2NZ
SIVM1
100
SIVG1
100
SIVV1
C
PRANK-Codon-CD
66
HV1B1
72
HV1J3
39
HV1JRFL
39
HV1C4
71
HV1OY
98
HV1A2
97
HV1RH
100
HV1EL
98
HV1ND
HV1MA
SIVCZ
76
HV2G1
99
HV2D1
47
HV2BE
99
HV2CA
100
HV2NZ
SIVM1
100
SIVG1
99
SIVV1
D
PRANK-Codon-AS
67
HV1B1
69
HV1J3
32
HV1JRFL
32
HV1C4
70
HV1OY
99
HV1A2
99
HV1RH
100
HV1EL
98
HV1ND
HV1MA
SIVCZ
65
HV2G1
100
HV2D1
28
HV2BE
99
HV2CA
100
HV2NZ
SIVM1
100
SIVG1
99
SIVV1
E
CAUSA-CD
42
HV1B1
27
HV1J3
34
HV1JRFL
20
HV1C4
58
HV1OY
99
HV1A2
91
HV1RH
100
HV1EL
96
HV1ND
HV1MA
SIVCZ
73
HV2G1
100
HV2D1
28
HV2BE
99
HV2CA
100
HV2NZ
SIVM1
100
SIVG1
100
SIVV1
F
CAUSA-AS
55
HV1B1
52
HV1J3
33
HV1JRFL
20
HV1C4
62
HV1OY
100
HV1A2
99
HV1RH
100
HV1EL
98
HV1ND
HV1MA
SIVCZ
69
HV2G1
99
HV2D1
16
HV2BE
99
HV2CA
100
HV2NZ
SIVM1
100
SIVG1
100
SIVV1