Accurate Reconstruction of Molecular Phylogenies for Proteins Using Codon and Amino Acid Unified Sequence Alignments (CAUSA)

Xiaolong Wang[1*], Yu Fu[1], Yue Zhao[1], Qi Wang[1], Chandra Sekhar Pedamallu[2], Shuang-yong Xu[3], Yingbo Niu[1]

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

2. The Broad Institute, Cambridge, MA 02142, USA and Dana-Farber Cancer Institute (DFCI), Cambridge, Boston, MA 0193802215, USA

3. New England Biolabs, Inc., Ipswich, MA 01938, USA

* To whom correspondence should be addressed, E-mail: ouqd@hotmail.com

ABSTRACT

Based on the molecular clock hypothesis, molecular phylogenies have been widely used for inferring evolutionary history of organisms and individual genes. Traditionally, alignments and phylogeny trees of proteins and their coding DNA sequences are constructed separately, thus often different conclusions were drawn. Here we present a new strategy for sequence alignment and phylogenetic tree reconstruction -- codon and amino acid unified sequence alignment (CAUSA). We demonstrated that CAUSA improves both the accuracy of multiple sequence alignments and phylogenetic trees by solving a variety of molecular evolutionary problems in virus, bacterial and mammals. Our results support the hypothesis that *the molecular clock for proteins has two pointers* existing separately in DNA and protein sequences. It is more accurate to read the molecular clock by combination (additive) of these two pointers.

\body

1. INTRODUCTION

In 1962, Emile Zuckerkandl and Linus Pauling first noticed that the number of amino acid differences in a specific protein was approximately constant over time, and lineages, which predicted a *molecular clock* [*1*]. In 1968, Motoo Kimura developed *the neutral theory of molecular evolution* [*2*]. In the early 1980s, Masatoshi Nei and his students initiated the study of inference of phylogenetic trees based on molecular distance data [*3-4*]. Later, in 1985, they developed the neighbor-joining and minimum-evolution methods of tree inference [*5*]. At present, the use of phylogenetic trees based on molecular clock to determine the classification of organisms or to study variation in proteins has been an important tool in molecular genetics, such as establishing the dates of speciation events, the divergence of living taxa and the formation of the phylogenetic trees [*6-7*].

However, phylogenetic trees based on single, or small numbers of, genes can differ from one another for statistical and evolutionary reasons, such as differences in evolutionary rates, convergent evolution, horizontal gene transfer, etc. For example, Wray [*8*] concluded from the analysis of seven genes that the divergence of protostomes and deuterostomes occurred nearly twice as early as the Cambrian, about 1,200 Million years ago (Mya), and that chordates diverged from the echinoderms about 1,000 Mya. However, Ayala [*9*] studied the origin of the metazoan phyla and confirm paleontological estimates by analyzing 18 proteins and suggested that the divergence of protostomes and deuterostomes occurred in the late Neoproterozoic, around 544–700 Mya.

In fact, in the original study [*3*], Nei pointed out that any tree-making method is likely to make errors in obtaining the correct topology with a high probability, unless all branch lengths of the true tree are sufficiently long. Recently, systematic biases were observed in simulated sequence analyses: Whelan [*10*] reported that the genetic code can cause systematic bias even in simple phylogenetic models; and Revell [*11*]

demonstrated that underparameterized model of sequence evolution leads to bias in the estimation of diversification rates from molecular phylogenies.

In recent years, due to increasingly wider availability of sequence data, it has been able to reveal functional and structural changes leading to genetic differences among different species, and provide accurate reconstruction of evolutionary histories of related genes, proteins and genomes. Evolutionary, structural or functional studies require accurate multiple sequence alignments (MSA), *i.e*. the correct identification of homologous nucleotides or amino acids (aa), and the accurate positioning of gaps indicating insertions and deletions (indels). However, present MSA and phylogeny reconstruction methods are not perfect, sometimes producing systematic bias, leading to subsequent misinterpretation of evolutionary or structural information.

CLUSTAL W [*12*] is by far the most widely used MSA tool. CLUSTAL W build a multiple alignment from pairwise alignments, performed in order of decreasing relatedness according to a guide tree using progressive multiple sequence alignment algorithm. Although there are quite a few new MSA tools, such as MUSCLE [*13*], MAFFT [*14*] and T-coffee [*15*], different method often lead to drastically different conclusions in sequence alignment and phylogenetic tree on a same set of sequence data, and support entirely different mechanisms driving evolutionary and structural changes [*16*]. In addition, alignments created by a computer program often require fine adjustments made by human visual inspection [*9*], which is cumbersome and a potential source of errors.

Moreover, traditionally multiple sequence alignments and molecular phylogenetic trees of protein and their coding DNA sequences are built separately, resulting in often very different conclusions. Here we present a new strategy for MSA and molecular phylogenetic tree reconstruction -- *codon and amino acid unified sequence alignment* (CAUSA), which improves the accuracy of both MSA and tree by uniting DNA and protein sequences and aligning them simultaneously in a unified fashion. We demonstrated the utility of CAUSA by constructing a variety of molecular evolutionary trees in virus, bacteria, and mammals.

2. RESULTS AND DISCUSSIONS

2.1 CAUSA improves the accuracy of MSA

As shown in Fig. 2A and S1A, a traditional protein alignment of HIV *env* aligned by CLUSTAL W shows that part of the variable (V_2) region has a high rate of substitutions. Löytynoja and Goldman [*16*] used this alignment as a typical example to show that traditional alignment tools incorrectly squeezed similar, but distinct, inserted sequences between two conserved blocks. They pointed out that this problem is actually caused by repeated penalizing gap-opening [*17*], but cannot be avoided by reducing gap-opening penalties, since it will result in 'gappy' alignments. Alignments given by other tools, such as MAFFT (Fig. S1B), MUSCLE (Fig. S1C) and T-coffee (Fig. S1D), are improved to some extent, but the problem of mismatching distinct insertions still exists.

In order to solve this problem, Löytynoja and Goldman [*16-17*] developed the phylogeny-aware (PRANK) method that "flags" the gaps introduced in earlier alignment steps, and so that distinct insertions are kept separate even when they occur at exactly the same position. As shown in Fig. S1E, using PRANK they identified several 'distinct' insertions. At the same time, however, they ignored homologies and similarities among these inserted sequences. They pointed out that inserted sequences are not descendants of any ancestral characters, and should not be aligned with anything [*16*]. However, if an inserted sequence is homologous to other sequences, it is possible that it is an ancestor of the later ones. Recently, Dessimoz and Gil [*18*] reported that phylogenetic assessment of alignments reveals neglected tree signal in gaps, present in the variable region, carried substantial phylogenetic signal, but are poorly exploited by most alignment and tree building programs (including PRANK). Therefore, proper alignment of insertion sequences potentially has a serious impact on downstream phylogenetic analysis, so similarities and homologies among them should not be simply ignored.

This problem can be solved by this CAUSA method that constructs unified alignments. In a unified alignment, 4-turples of codon-aa and gaps show every

detailed mutation event, such as insertions, deletions, synonymous and non-synonymous base substitutions. In a conserved region, the rate of synonymous substitutions is dominantly higher than that of non-synonymous ones. In a variable region, however, the rate of non-synonymous substitutions is higher than in the conserved regions. These unified codon-aa 4-turples, together with the 64-color views, make it much easier to distinguish substitutions from indels. In the protein alignment of *env*, *e. g.*, some inserted sequences seems to be 'homologous' to each other in the protein view (Fig. S2A), but significant differences were shown in the unified view (Fig. S2B) or DNA view (Fig. S2C) back-translated from the protein view. In the unified alignment (Fig. S2D), however, not only these insertions were correctly identified, but also homologies among them were clearly shown. In addition, as indicated by solid arrows in Fig. S1A through S1G, more accurate alignment for conserved residues was given in this variable region. Obviously, CAUSA has a much more powerful ability to distinguish substitutions from indels.

2.2 CAUSA improves phylogenetic analyses of virus genomes

Traditional progressive algorithms perform heuristics pairwise alignments at the branching points of a guide phylogenetic tree approximating the evolutionary history of DNA or protein sequences. However, different tools often give different alignment and phylogenetic trees. As shown in Fig. S3A to S3E, *e. g.*, the phylogenetic trees of HIV *env* given by CLUSTAL W, MAFFT, T-coffee, MUSCLE and PRANK are all varied greatly, and the tree inferred from the unified alignment suggests another different evolutionary process (Fig. S3F).

In order to compare and evaluate the accuracy of these different alignment algorithms and phylogenetic trees, we build alignments and trees for two HIV genes, *env* and *gag,* respectively using protein alignments, codon-based DNA alignments and unified alignments. Through systematically examining phylogenetic trees of SIVs in different genomic regions, it was concluded that the chimpanzee SIV (SIVcpz) is mosaic: the 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 [*19*]. The mosaic structure of SIVcpz requires that a chimpanzee were infected with two different monkey viruses and these recombined. It is likely that the dual infection occurred in a chimpanzee, since chimpanzees hunt and eat these two kinds of monkeys [*19*].

Since HIV was originated from SIVcpz, it is therefore interesting to ask whether such kind of dual infection and recombination had also happened in HIV. Bootscan analysis, which breaks HIV genomes into small sections and analyzes each section independently, has been used to identify areas of recombination within HIV genomes [*20*]. However, the apparent phylogenetic incongruence at different regions of the genome that was taken as evidence of recombination was shown to be not statistically significant [*21*]. A more likely explanation for the differences in the evolutionary rates across the genome is that different regions of the genome were under different selective pressures [*21*].

As shown in Fig. 3A, phylogenetic trees for *env* and *gag* genes constructed from protein, codon and unified alignments are all different from each other. Moreover, since the protein trees are different between *env* and *gag*, it seems that some of the HIV genomes, such as HV1J3, HV1B1, HB1A2, HV2BE and HV2G1, are recombinant forms. However, the unified trees of the two genes are fully consistent (Fig. 3B), suggesting that different regions of these HIV genomes had a same evolutionary process. Therefore, it seems that dual infection and recombination had never happened among these HIV strains since isolated from SIVcpz. We believe that these unified trees are more reliable than protein trees, not only because they are fully consistent between the two HIV genes, but also because they have higher Bootstrap percentages. In fact, they also have a biologically more significant theoretical basis, as described in the Supplementary material.

Codon alignment is an alignment model that takes into accounts both DNA and protein sequences [*22-25*], and gives a DNA alignment and a translated protein alignment. However, the DNA tree (Fig. 3C) and protein tree (Fig. 3D) for *env* and *gag*, inferred respectively from codon-based DNA alignment and protein alignment, are different from each other, and are even more inconsistent between these two genes

when compared with the CLUSTAL W protein trees. We compared CAUSA with CAT by back-translating codon-based protein alignments into unified alignments. In highly similar sequences, *env* for example, the codon alignment (Fig. 2C) and the unified alignment (Fig. 2B) are highly consistent, while CAUSA obviously outperforms CAT in more diverged sequences: for unknown reason, CAT misaligns conserved residues (indicated by red boxes in Fig. 2C) often in the variable region, which is obviously the cause of inconsistencies in the phylogenetic trees.

In addition, in hepatitis B virus (HBV) the unified alignment of the surface antigen (HBsAg) suggested a spreading path that is more plausible than those inferred from protein or codon alignments considering the geographical distribution of these HBV strains (Table S3 and Fig. S4). Artifacts of traditional protein- or DNA-only alignments may cause misinterpretation of a flawed mosaic genome structure or a false spreading path of virus, and inappropriate attribution of recombinant origins to divergent sequences obscures the evolutionary and epidemiology properties of viruses [*21*]. CAUSA provides an accurate tool that can prevent many, if not all, of these types of errors, and confirms the recombination forms and spreading paths of viruses with higher confidence. In addition, as described in the Supplementary material, CAUSA alignments and trees help better interpreting the mutation events happened in the evolutionary process.

2.3 CAUSA alignments improves phylogenetic analyses of bacterial proteins

We constructed unified alignments and phylogenetic trees for evolutionarily conserved and functionally important bacterial proteins, such as DNA/RNA polymerases, DNA topoisomerase and helicase, respectively using CLUSTAL W, CAT and CAUSA and compared them with a multi-gene phylogenetic tree derived from the PathoSystems Resource Integration Center (PATRIC) [*26*]. When compared with protein and DNA trees in more than thirty bacterial proteins (Table S4), we concluded that unified trees are significantly more consistent with the multi-gene phylogenetic tree, suggesting that unified alignments are more accurate, and superior in phylogeny analysis, than protein or codon alignments. In DNA topoisomerase III, *e.*

g., *B. pertussis Tohama I*, a strict human pathogen and the primary etiologic agent of whooping cough, is grouped as a descendant both in the protein trees (Fig. 4A) and the DNA trees (Fig. 4B), but an ancestor of the other strains in the unified tree (Fig. 4C). The unified tree is fully consistent with the multi-gene phylogenetic tree of *Bordetella,* a group of *Proteobacteria* (Fig. 4D). The multi-gene phylogenetic tree is considered to be very reliable, since it has been reported that *B. pertussis* is one of independent derivatives of *B. bronchiseptica*-like ancestors, which infects smaller mammals (cats, dogs, rabbits, etc.), but not human [*27*].

In addition, unified alignments and trees are useful for the evolutionary analysis of bacterial restriction enzymes. Two typical examples are given in the supplementary materials: the unified tree for BamHI homologs is different from the protein tree and the DNA tree, and with higher Bootstrap percentages (Fig. S5); and that of SauUSI homologs, a group of Type IV modification-dependent restriction enzymes that were recently discovered in *Staphylococcus aureus subsp. aureus* USA300 [*28*], is more consistent with the PATRIC multi-gene phylogenetic tree (Fig. S6). Bacterial restriction-modification (R-M) systems encoding enzymes for DNA methylation and endonuclease activity are subject to rapid evolution and are sometimes associated with mobile genetic elements, such as transposon, phage and plasmid. Closely-related isoschizomers found in diverse bacterial species suggest that R-M systems can be acquired by horizontal gene transfer (HGT) [*29*]. More accurate alignments and trees distinguish real horizontally-transferred genes from flaws, thus helps better understanding of evolutionary relationship among the few thousand R-M systems.

2.4 Phylogeny-based testing alignment and tree accuracy in mammals

Recently, Dessimoz and Gil [*18*] reported phylogeny-based testing of MSA accuracy using large and representative samples of real biological data. According to Fitch's definition of orthology [*30*], trees inferred from orthologs are expected to have the same topology as the underlying species. Thus, if a particular method produces alignments that result in trees more frequently congruent with the phylogeny of the species, it is likely to be more accurate [*18*]. Following this principle, we compared trees constructed from different alignments for more than fifty orthologous protein

families in human and mammalian species whose phylogeny, *Tree of Life* (TOL) [*31*], is undisputed.

Highly conserved protein families, such as cytochrome oxidases and histones, have long been used as benchmark for testing alignment algorithms and phylogeny reconstruction methods. Recently, a mitochondrial-gene-encoded protein, cytochrome oxidase subunit I (COXI), has been used as a standard DNA barcode for animal species identification [*32-33*]. The unified alignment of COXI (Fig. 5A) is totally consistent with protein or DNA alignments created by any other aligners: there is no gap throughout the whole alignment. However, the protein tree (Fig. 5B), the DNA tree (Fig. 5C), and the unified tree (Fig. 5D) are all different from each other, and the unified tree is, however, most consistent with TOL (Fig. 5E).

Similar results were observed in most of the closely-related proteins (pairwise similarity>85%) we tested, including twelve mitochondrial-gene-encoded (Table S5) and forty-two nuclear-gene-encoded (Table S6) proteins that were randomly selected from human and mammalian animals (Table S7). Two examples of nuclear-gene-encoded proteins, histone H3b (Fig. S7) and Doublesex- and mab-3 related transcription factor 1 (DMRT1) (Fig. S8), are shown in the supplementary material. Histone H3b is highly conserved, and three types of alignments are fully consistent; while *dmrt1* has a variable region, in which different types of alignments differ greatly. However, in both of these two proteins unified phylogenetic trees are more consistent with TOL than their corresponding protein or DNA trees. In addition, as shown in the statistics of *dmrt1* alignment (Table S8), the average number of gaps in the unified alignment is close to that in the other alignments, but the average number of indels is two-fold larger, and therefore the average length of indels is a half smaller, than those of the other alignments. Obviously the unified alignment prefers supporting short indels rather than forcing them into long-running gaps.

Pairwise comparisons (unpaired t-test) were used to assess differences of numbers of branches that are consistent with TOL for each tree, we found that unified trees are significantly more consistent with TOL when compared with other types of phylogenetic trees (Table 1), including protein trees inferred from protein alignments

by CLUSTAL W, MAFFT, MUSCLE, T-coffee, and DNA trees inferred from codon alignments by CAT. Overall, we observed that codon-based DNA alignments are more accurate than amino-acid alignments, while the best alignments and trees are obtained from unified codon-aa alignments.

2.5 Unified trees are more accurate than DNA and protein trees on simulated sequences

Usually, a MSA algorithm should be tested and compared with other methods on a set of hand-curated alignment benchmarks. However, current existing benchmark datasets, such as BAliBASE [*34*], contains only protein alignments aligned and adjusted at the amino acid level. Conceivably, it is unfair, and meaningless, to assess unified alignments using a benchmark based on amino acid alignments. Therefore, in addition to real biological data, simulated protein-coding DNA sequences were used to assess the unified alignment and tree reconstruction method.

As shown in Fig. S9, protein-coding DNA sequences were simulated using program *Recodon* v1.6.0 [*35*] with parameters pre-determined from the above human and mammalian sequence data. Phylogeny trees constructed by each method were assessed by counting the number of correct branches compared with the *true tree* given by the program. In sequences simulated by Recodon, as shown in Table S9 and S10, unified trees are significantly better than protein trees, and slightly better than DNA trees. Both DNA and unified trees for these simulated sequences are highly accurate. However, in the above human and mammalian data much higher rate of inconsistencies was observed in all three kinds of trees, and unified trees are significantly more consistent with TOL when compared with the other two types of trees (Table S10). Dessimoz and Gil pointed out that simulated sequences strongly depend on the choice of codon model used to generate the data, as most biological processes are difficult to model realistically, so that a simulation will never fully capture the complexity of real biological data [*18*]. The result observed in simulated sequences is somewhat different from those in the real biological data. Nevertheless, unified trees are more reliable than protein and DNA trees even in such kind of sequences simulated using a simple model, which means that a systematic bias is

present in traditional protein and DNA trees.

In addition, as shown in both real (Table 1) and simulated data (Table S10), unified trees have significantly higher average Bootstrap percentages than protein trees, difference of average Bootstrap values between unified trees and DNA trees is, however, not statistically significant. This suggested that the higher average Bootstrap percentages in unified trees are not flawed due to duplicated codon-aa sequence information, since otherwise Bootstrap values of unified trees should also be significantly higher than that of the DNA trees. In fact, duplicated information will in principle neither alter the topology of the phylogenetic trees, nor increase overall average Bootstrap percentage, since Bootstrap testing itself is a statistic method that uses duplicated data resampling. Therefore, the higher average Bootstrap value in unified trees are truly because of the uniting of DNA and protein sequences, which contain duplicated but, to some extent, different information; and it is the different but not the duplicated portion of DNA and protein sequence information that result in the different topologies and Bootstrap values among unified, DNA and protein trees. Therefore, we concluded that, by integrating information buried separately in DNA and protein sequences, CAUSA allows more accurate, and more confident, reconstruction of molecular phylogenies for closely-related proteins.

Since sequences simulated by Recodon do not incorporate indels, there is no need for sequence alignment and no way to evaluate alignment accuracy. We further tested unified alignments in coding DNA sequences with indels that were simulated by programs *indel-seq-gen* v 2.1.03 [*36*]. As shown in Fig. S10, three types of alignments and trees are almost always totally consistent and almost all completely correct in sequences with small or moderate proportions of indels, but in sequences with high proportions of indels, three types of alignments and trees are all highly error-prone, and too many stop codons are generated. Although we can clearly see that unified alignments outperform protein and codon alignments in real data, mainly in the variable region, we did not observe the same benefit in the sequences simulated by indel-seq-gen, which is presumably due to current insertion-deletion simulation models are known to be insufficient [*36*].

2.6. Comparing methods for aligning protein-coding DNA sequences

Due to the small size of the alphabet of nucleotide bases, alignment of DNA sequences is inherently difficult: even two completely unrelated DNA sequences will display ~25% identity over their entire length and it is often possible to find extended local alignments where >50% of nucleotides are identical [*37*]. This makes it difficult to distinguish true homology from random similarity. Proteins are built from 20 amino acids while DNA contains only four bases, so that the 'signal-to-noise ratio' in protein sequence alignments is much better than that of DNA sequences [*37*]. Besides this advantage in theoretical information-content, protein alignments also benefit from amino acid substitution matrices, such as PAM [*38*], BLOSUM [*39*] and Gonnet [*40*] series. These matrices contain empirically derived scores for each possible amino acid substitution and provide a rational basis for aligning amino acids.

In addition, due to overall higher rates of synonymous over nonsynonymous substitutions [*41-42*], it has been believed that the phylogenetic signal disappears more rapidly from DNA sequences than from their encoded proteins, and therefore preferable to align protein coding DNA sequences at amino acid level [*37*]. However, some important information carried by DNA sequences, such as synonymous substitutions and frame-shift mutations, get lost after they were translated into amino acid sequences, makes the resulting alignments and trees somewhat inaccurate. Given the substantial evolutionary time separating the animal phyla, for example, the statistical noise associated with the substitution process leads to a high probability that phylogenetic trees based on different proteins will yield different topologies, so that inferences based on single genes can potentially be very misleading [*43*]. Multi-gene phylogenetic trees have been therefore widely used in phylogenies analyses of various organisms. However, the problem of reconstructing phylogenetic trees for individual protein has not been sufficiently addressed.

Several strategies have been developed to deal with this problem. The first is to construct a DNA alignment by back-translating a protein alignment, such as RevTrans [*37*]. The second method, Hein's COMBAT [*22-23*], combines a DNA alignment and a protein alignment into a combined alignment. And the third is to construct a codon

alignment that takes into account both DNA and protein information, and attempts to minimize the total amount of mutation at both DNA and protein levels [*24-25*]. However, phylogenetic trees for DNA and protein sequences were all constructed separately, thus often different conclusions were drawn. Here by using unified DNA-protein scoring matrices, CAUSA aligns protein and their encoding DNA sequences simultaneously in a single alignment. The *position effect* of the arrangement of codon-aa 4-turples, together with the high-penalties (-99) that naturally prevents mismatches between aa and bases, helps better aligning both DNA and protein sequences, like writing every English word followed by its translation in Chinese, helps readers to understand both languages more easily.

2.7 Conclusion

Multiple sequence alignment is the starting point of studies in molecular genetics and genomics such as reconstruction of phylogeny history, protein structure modeling and functional analyses [*45-46*]. Our analysis shows that errors in traditional protein or DNA alignments and phylogenetic trees may lead to inconsistency and errors in evolutionary and comparative studies even in closely-related proteins. However, it is not that the progressive algorithm itself is defective. Rather, accurate alignment and phylogeny analysis requires that information carried by proteins and their coding DNA sequences to be integrated and exploited in a unified manner.

In addition, unified trees are more consistent with evolution histories than protein and DNA trees in various species tested, supporting the hypothesis that *the molecular clock for proteins has two pointers,* as schematically shown in Fig. S11, existing in DNA and protein sequences that are undergoing convergent evolution; and it is more accurate to read the molecular clock by the additive of these two pointers, since the ticking rates of them are sometimes consistent, sometimes different. Combining information buried separately in DNA and protein sequences, CAUSA allows homologous sites to be aligned more accurately, overcomes the problems commonly exist in conventional DNA or protein alignment, thus gives a more accurate picture for protein evolution, and raised the question of how alignment and phylogeny of non-coding DNA and RNA sequences could be inferred accurately.

3. Materials and Methods

3.1 Protein coding DNA sequences and online resources

Proteins and their orthologs in representative species, including virus, bacteria, mammals and human, were derived from online protein family databases, including pFAM, TreeFam and CDD. Their coding DNA sequences (CDSs) were retrieved from GenBank or EMBL nucleic acid databases using Ensembl, Homologene or NCBI BLAST tools.

3.2 Converting and aligning CDSs by CAUSA

As shown in Fig. 1, using an in-house developed computer program, CAUSA, protein-coding DNA sequences of interest are translated into amino acids and converted into codon and amino acid unified sequences (CAUSs), in which every triplet codon is immediately followed by the one-letter code of its encoded amino acid. In CAUSs, every information unit consists of a triplet codon followed by its encoded amino acids, which are called *codon-aa 4-turples* and shown in 64-color views. CAUSs were then aligned by calling CLUSTAL W using a c*ombined DNA-Protein (CDP) scoring matrix*, such as *CDP-Gon250 matrix* (Table S1), and a set of user-defined settings (Table S2). The principle and implementation of the CAUSA algorithm are described in details in the Material and Method section in the supplementary material. The CAUSA software are released as Open Source and downloadable free of charge from website www.dnapluspro.com.

3.3 DNA, Protein, and Codon Alignments

Conventional DNA or protein alignments are constructed using CLUSTALW v. 2.0.12, MAFFT v. 5.861, MUSCLE v. 3.6, T-COFFEE v. 3.93 and PRANK. Codon Alignments were constructed using an online codon alignment tool (CAT) provided by the HIV database (http://www.hiv.lanl.gov/), which is maintained at Los Alamos National Security, LLC (LANS) and supported by the NIH and DOE. All programs were run with default settings.

3.4 Unified and DNA alignments back-translated from protein alignments

As shown in Fig. 2A, S1A to S1F, and S2A, using CAUSA software, protein alignments can be back-translated into DNA alignments and unified alignments, so that a protein alignment aligned by other aligners can be compared with a corresponding unified and DNA alignment in a unified view.

3.5 Phylogenetic trees

Phylogenetic trees for individual protein coding genes were constructed respectively from protein alignments, codon-based DNA alignments and unified alignments. The evolutionary history was inferred using the Neighbor-Joining method [*5*]. Phylogenetic trees were drawn using MEGA v5.05 [*6*]. The percentages of replicate trees in which the associated taxa clustered together in the Bootstrap test (1000 replicates) are shown next to the branches [*7*]. The evolutionary distances were computed using the *p*-distance. All sites containing gaps and missing data were eliminated (Complete deletion option). Multi-gene phylogenetic trees for bacteria were inferred from the PathoSystems Resource Integration Center (PATRIC) (http://patricbrc.vbi.vt.edu). Phylogeny trees of mammalian species were derived from the *Tree of Life Web Project* (http://tolweb.org).

Acknowledgements

This study was supported by China National Science Foundation through grant 81072567 and Shandong Provincial Natural Science Foundation through grant ZR2010HM056. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Zuckerkandl E, Pauling LB (1962) Molecular disease, evolution, and genetic heterogeneity In Kasha, M and Pullman, B (editors) *Horizons in Biochemistry.* Academic Press, New York pp 189–225.

2. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217 (5129): 624–626.

3. Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Mol. Evol.* 18: 387-404.

4. Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol.* Evol. 19: 153-170.

5. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

6. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599.

7. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783-791.

8. Wray GA, Levinton JS, Shapiro LH (1996) Molecular evidence for deep Precambrian divergences among Metazoan phyla. *Science* 274, 568–573.

9. Ayala FJ, Rzhetsky A, Ayala FJ (1998) Origin of the metazoan phyla: Molecular clocks confirm paleontological estimates. *Proc. Natl. Acad. Sci. USA* 95: 606–611.

10. Whelan S (2008) The genetic code can cause systematic bias in simple phylogenetic models. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 363(1512): 4003-4011.

11. Liam J, Revell, Luke J, Harmon, Glor ER (2005) Under-parameterized Model of Sequence Evolution Leads to Bias in the Estimation of Diversification Rates from Molecular Phylogenies. *Syst. Biol.* 54(6): 973-983.

12. Thompson JD, Higgins DG, Gibson TJ (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.

13. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30(14): 3059–3066.

14. Edgar CR (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5): 1792–1797.

15. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302(1): 205-217.

16. Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* 320 (5883): 1632-1635.

17. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA* 102(30): 10557–10562.

18. Dessimoz C, Gil M (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.* 11(4): R37.

19. 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.

20. Salminen MO, Carr JK, Burke DS, McCutchan FE (1995) Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res. Hum. Retrovir.* 11: 1423–1425.

21. Anderson JP, et al (2000) Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. *J. Virol.* 74(22): 10752-10765.

22. Hein J (1994) An algorithm combining DNA and protein alignment. *J. Theor. Biol.* 167: 169–174.

23. Hein J, Støvlbæk J (1996) Combined DNA and protein alignment. *Methods Enzymol.* 266: 402–418.

24. Hua Y, Jiang T, Wu B (1999) Aligning DNA Sequences to Minimize the Change in Protein. *J. Combinatorial Optimization* 3: 227-245.

25. Stocsits RR, Hofacker IL, Fried C, Stadler PF (2005) Multiple sequence alignments of partially coding nucleic acid sequences. *BMC Bioinformatics* 6: 160.

26. Snyder EE, *et al* (2007) PATRIC: The VBI PathoSystems Resource Integration Center. *Nucleic Acids Res.* 35 (Database issue) 401-406.

27. Parkhill J, *et al* (2003) Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat. Genet.* 35(1): 32-40.

28. Xu SY, Corvaglia AR, Chan SH, Zheng Y, Linder P (2011) A type IV modification-dependent restriction enzyme SauUSI from *Staphylococcus aureus subsp aureus* USA300. *Nucleic Acids Res.* 39(13):5597-5610.

29. Corvaglia AR, et al (2010) A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical *Staphylococcus aureus* strains. *Proc. Natl. Acad. Sci. U S A* 107(26): 11954- 11958.

30. Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19: 99-113.

31. Maddison DR, Schulz K-S (eds) (2007) The Tree of Life Web Project. Internet address: http: //tolweborg.

32. Hebert PDN, Ratnasingham S, DeWaard JR (2003) Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B. Biol. Sci.* 270: S596–S599.

33. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA Barcodes. *PLoS Biol.* 2(10): e312.

34. Thompson JD, Plewniak F, Poch O (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* 15(1): 87-88.

35. Arenas M, Posada D (2007) Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography. *BMC Bioinformatics* 8: 458.

36. Strope CL, Abel K, Scott SD, Moriyama EN (2009) Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol. Biol. Evol.* 26: 2581-2593.

37. Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* 31(13): 3537-3539.

38. Henikoff S, Henikoff, JG (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915–10919.

39. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G (2008) BLOSUM62 miscalculations improve search performance. *Nat. Biotech.* 26 (3): 274–275.

40. Gonnet GH, Cohen MA, Brenner SA (1992) Exhaustive matching of the entire protein sequence database. *Science* 256: 1443–1445.

41. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17: 32–43.

42. Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46: 409–418.

43. Lynch M (1999) The Age and Relationships of the Major Animal Phyla. *Evolution*. 53: 319-325.

44. Whelan S (2008) Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol. Biol. Evol.* 25, 1683–1694.

45. Notredame C (2007) Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Comput. Biol.* 3(8): e123.

46. Kemena C, Notredame C (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*. 25(19): 2455-2465.

Figure Legend

Fig. 1. The working flowchart of different strategies for aligning proteins and their coding DNA sequences. **(A)** Codon alignment; **(B)** Protein alignment; **(C)** CAUSA.

Fig. 2. Comparison unified views of different alignments of the variable (V_2) region of HIV gp120 protein (Env). **(A)** CLUSTAL W protein alignment; (**B**) CAUSA unified alignment; (**C**) CAT Codon alignment. HIV or SIV strains were derived from the seed alignment of Pfam gp120 protein family (pf00516). DNA and protein sequences are written respectively in lowercase and uppercase letters.

Fig. 3. Unified trees for *env* and *gag* suggest more consistent evolutionary process for different HIV genes. **(A)** The protein trees from CLUSTAL W. **(B)** The unified trees from CAUSA. **(C)** The DNA trees from codon alignments. **(D)** The protein trees from codon alignments.

Fig. 4. The unified tree of DNA topoisomerase III is fully consistent with multi-gene phylogenetic tree. **(A)** Protein tree. **(B)** DNA tree. **(C)** Unified tree. **(D)** Multi-gene phylogenetic tree for *Bordetella,* a group of *Proteobacteria,* inferred from PATRIC http://patricbrc.vbi.vt.edu/.

Fig. 5. Comparing alignments and phylogenetic trees for COXI of human and representative mammalian species. **(A)** The unified alignment of COXI; **(B)** The protein tree. **(C)** The DNA tree. **(D)** The unified tree. **(E)** A phylogeny tree of human and mammals inferred from *The Tree of Life Web Project* (http://tolweb.org).

Table 1. The t-test results of average number of correct branches and bootstrap percentages compared with Unified trees in human and mammalian species

Program	Tree type	Average Number of branches			Bootstrap percentage	
		Total	Correct	P-value	Average	P-value
ClustalW	Protein	12.11	6.85±2.46	2.15E-05**	54.34±14.43	2.56E-09**
Mafft	Protein	12.11	6.97±2.80	6.77E-04**	55.19±24.17	5.49E-05**
MUSCLE	Protein	12.11	6.74±2.80	1.58E-04**	55.10+24.36	7.00E-05**
T-coffee	Protein	12.11	6.81±2.85	3.27E-04**	55.07±24.07	4.89E-05**
CAT	DNA	12.11	7.50±1.99	0.0031**	61.52±8.95	0.0523*
CAUSA	Unified	12.11	8.04±1.87		62.55±8.60.13	

Note: Calculated from 12 mitochondrial- and 42 nuclear-gene-encoded proteins in human and 19 mammals, see Table S7 for details.

Fig. 1

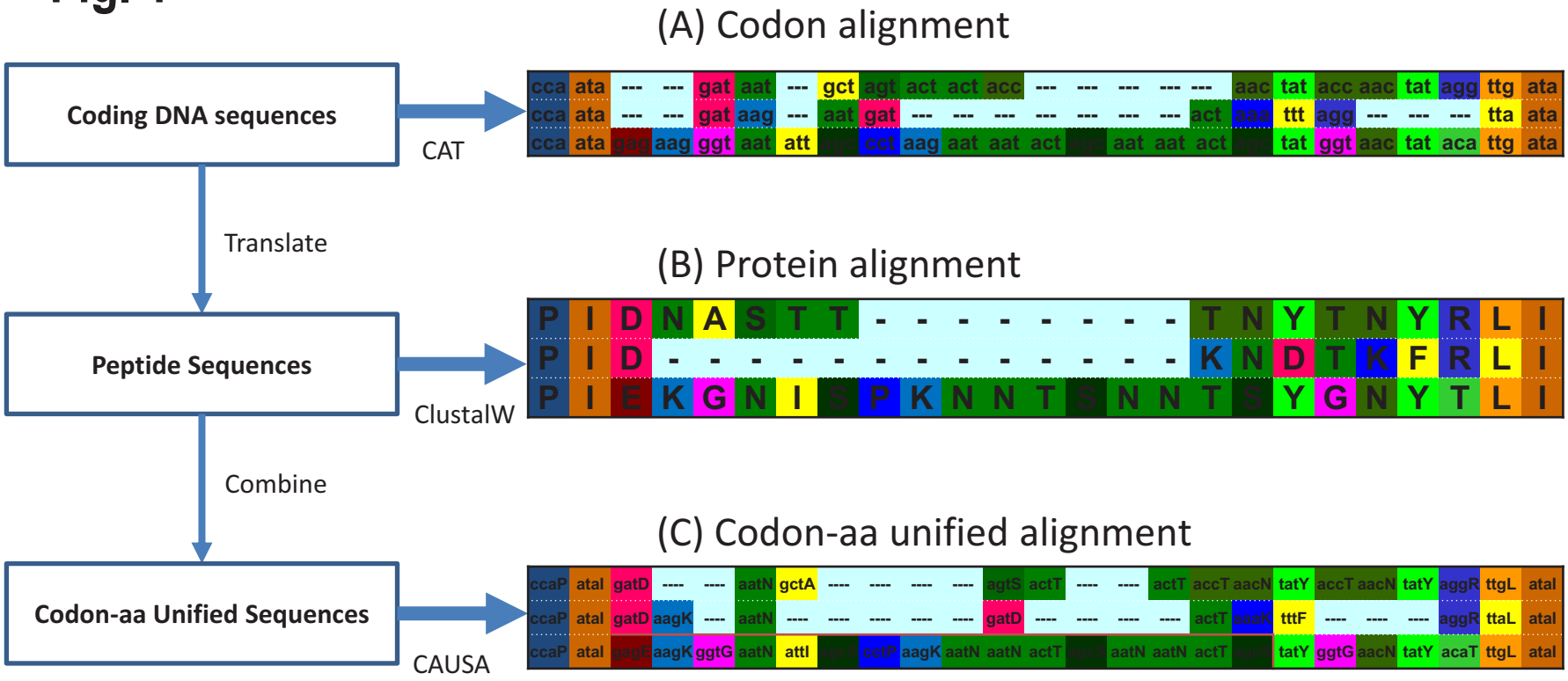
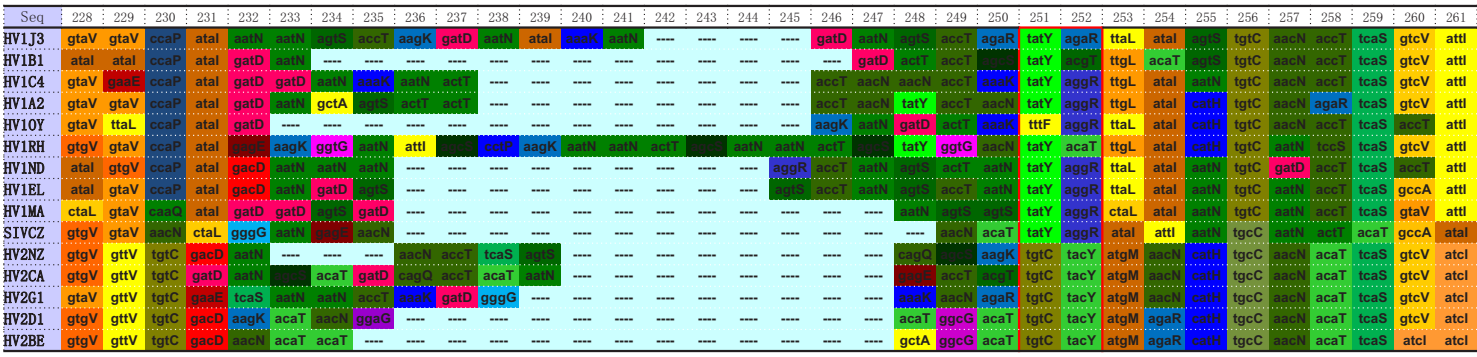
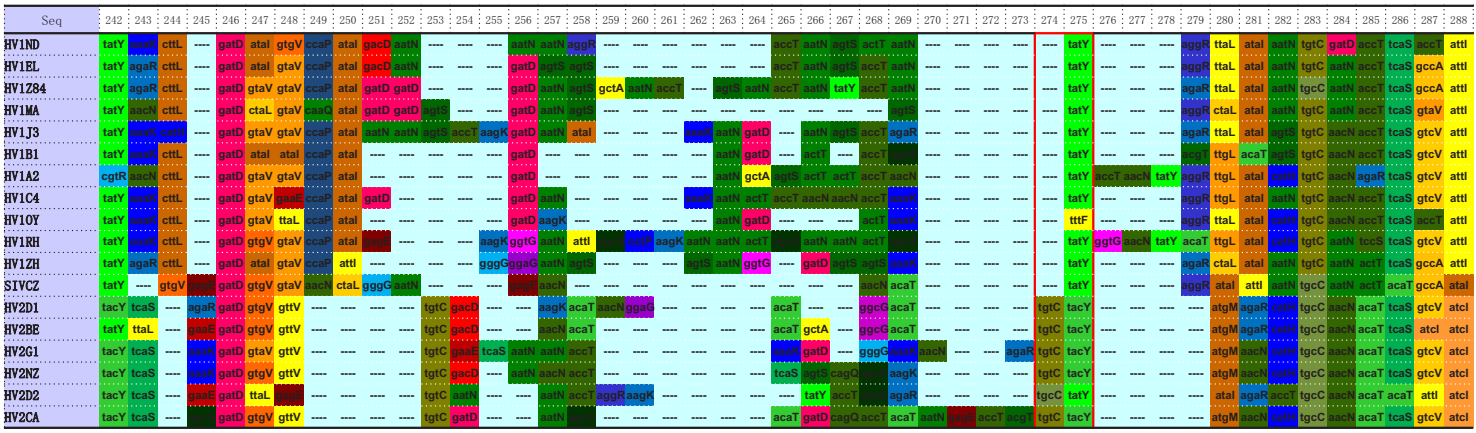


Fig. 2

A ClustalW



B CAUSA



C CAT

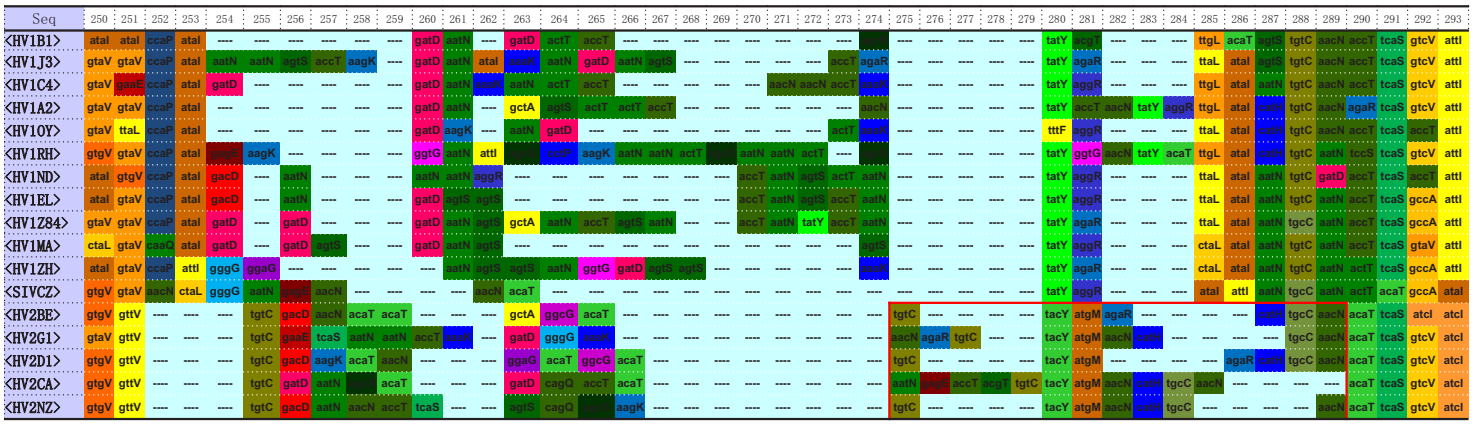
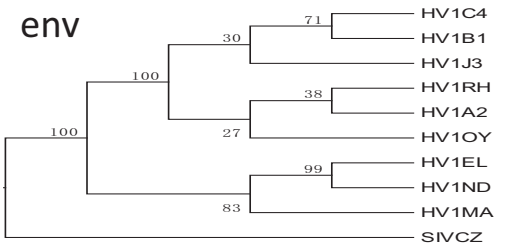
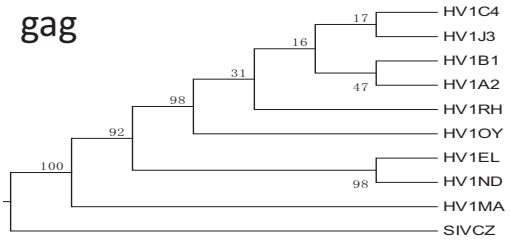


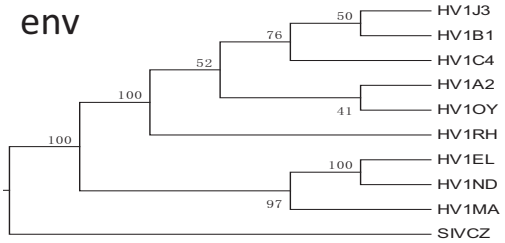
Fig. 3

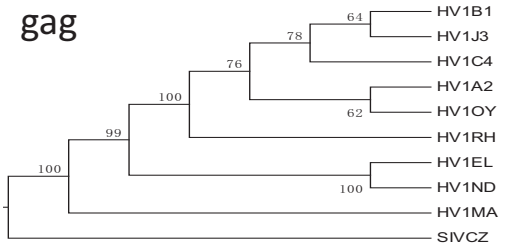
A
CLUSTALW
Protein tree



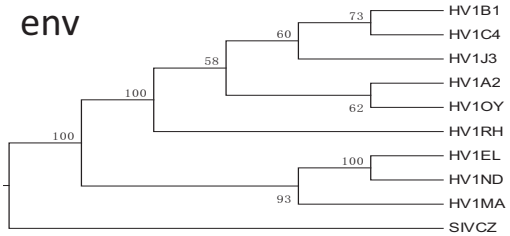


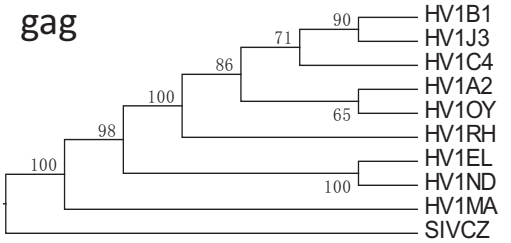
B
CAUSA
Unified tree



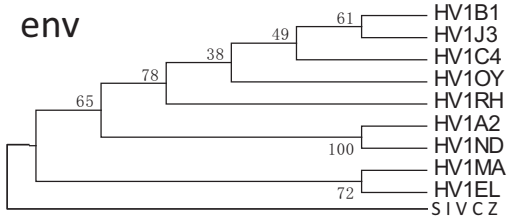


C
CAT
DNA tree





D
CAT
Protein tree



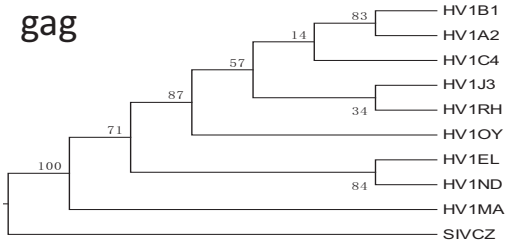
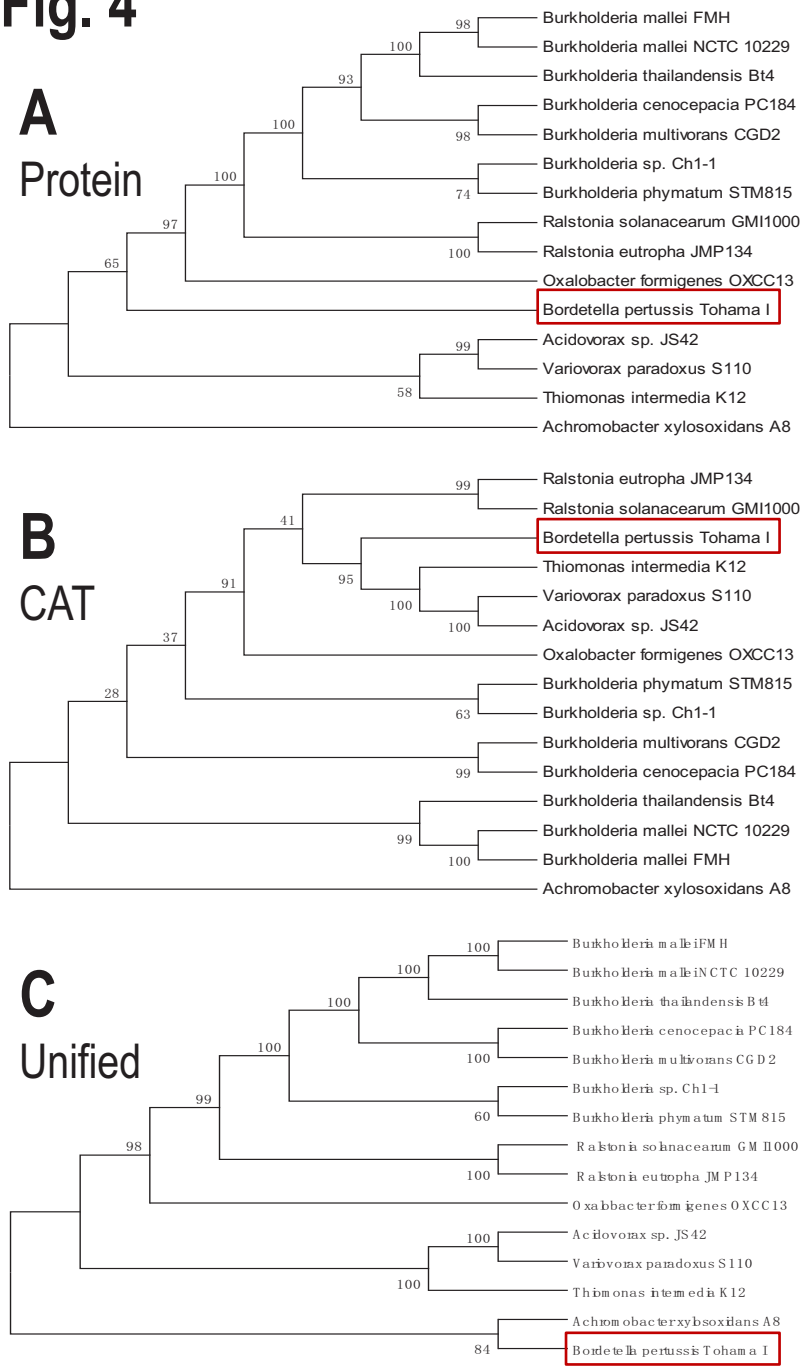
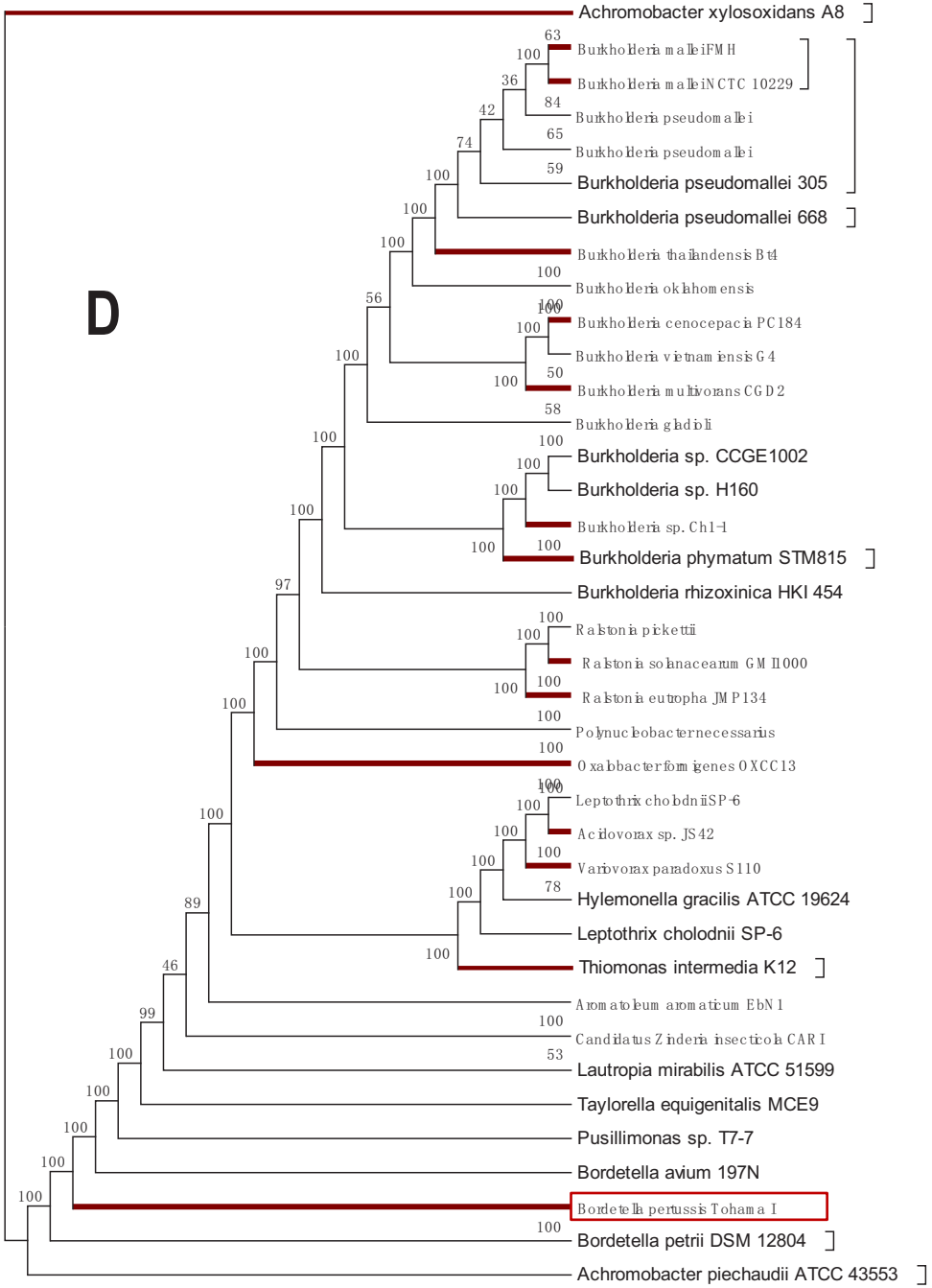
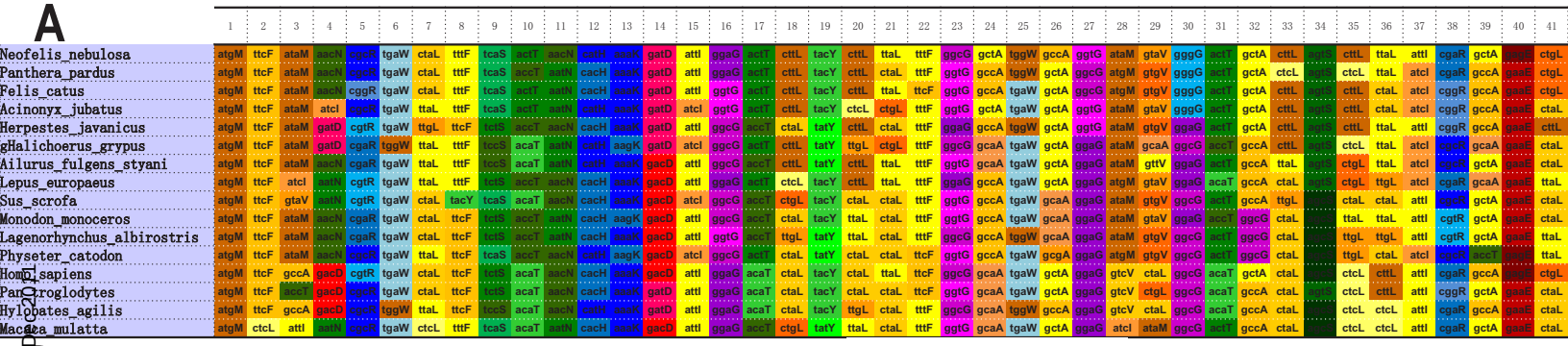
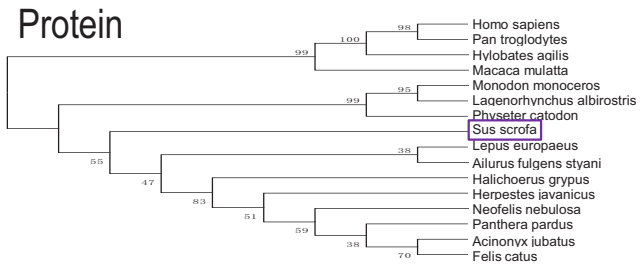


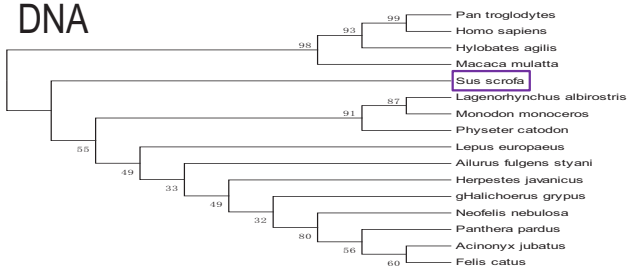
Fig. 4

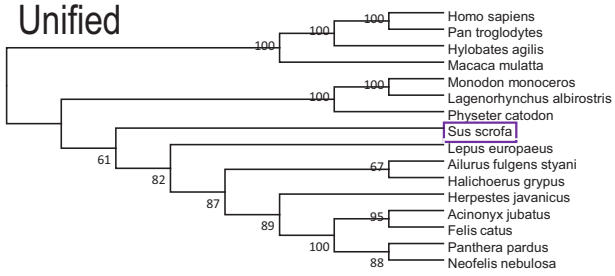












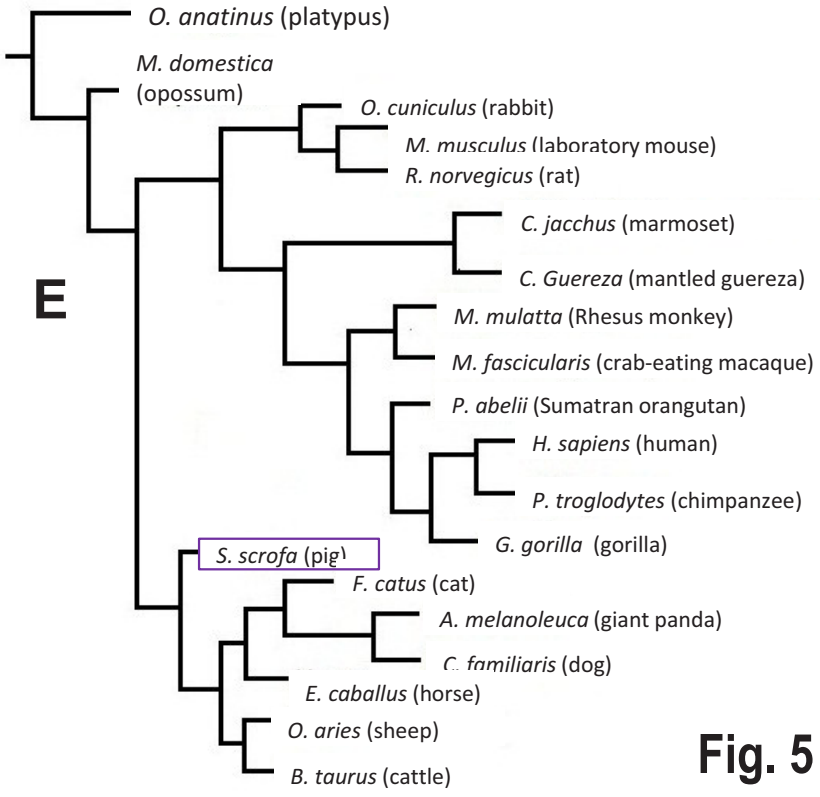


Fig. 5