Sequence Alignment in HIV Computational Analysis

Ana Abecasis[1,2,*], Anne-Mieke Vandamme[1,3], and Philippe Lemey[1,4]

[1] Laboratory for Clinical and Epidemiological Virology, REGA Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium
[2] Laboratory of Virology, Hospital de Egas Moniz, Centro Hospital de Lisboa Ocidental, Lisboa, Portugal
[3] Center for Bioinformatics, Instituto Nacional de Saúde Dr. Ricardo Jorge, Lisboa, Portugal
[4] Department of Zoology, University of Oxford, Oxford, UK
** Corresponding author, ana.abecasis@uz.kuleuven.ac.be*

Introduction

Aligning nucleotide or amino acid sequences is a very common procedure in HIV computational analysis. An accurate alignment is the first step in making a proper and correct analysis of HIV datasets. Sequence alignments are essential for phylogenetic analysis tracing the epidemiology of HIV, but also for interpretations of drug resistance and data mining efforts, where correctly positioning nucleotides or amino acids of different strains with respect to each other is pivotal.

Positional homology

If two sequences have a common ancestor, they are said to be homologous. By aligning them, we are inferring positional homology from statistically significant sequence similarity: any two sequences have some measurable similarity, but a statement of homology implies that this similarity is a specific result of common ancestry (47). Only when the common ancestry is recent enough, homology will still be reflected in a sufficient similarity, allowing an unambiguous alignment. When we align sequences, we are therefore looking for evidence that they have diverged from a common ancestor by evolutionary processes like selection and mutation (substitutions, insertions, and deletions) (9). Hence, the process of alignment is intimately related to inference of evolutionary relationships among sequences. In fact, the ideal alignment algorithm would allow us to co-estimate sequence alignment and phylogeny.

It is not surprising that the quality of an alignment will depend on the degree of sequence divergence, in particular the frequency of insertions and deletions (collectively referred to as indels) that have occurred. In HIV sequences, indels are frequently observed, even at relatively low divergence or over short evolutionary times. Figure 1a shows a short amino acid alignment of HIV envelope sequences sampled from a single host (positions 346 to 416 in gp120 according to HXB2 numbering). The sequences were obtained from plasma and different cellular populations at two different time points separated by approximately two years (51). Although this represents only a relatively short evolutionary time, the sequences cannot be unambiguously aligned in the hypervariable loop (V4). This example illustrates that any alignment procedure will generate an output for which the quality cannot be guaranteed. Obviously, alignment quality becomes even more problematic at greater evolutionary scales. For other viruses like Hepatitis C (HCV), indels may be observed less frequently relative to nucleotide substitutions, making the alignment process more straightforward.

Any two sequences can be fed into an alignment algorithm, and an alignment will be provided. When such an alignment requires many indels, and the aligned nucleotides do not seem to be unambiguously aligned, such an alignment cannot be considered successful in achieving positional homology. Some general guidance in assessing alignment quality can be found in the overall sequence similarity: the “twilight zone” between unambiguous and ambiguous alignment is considered to lie between 50% and 60% sequence identity for nucleotide sequences (14), and between 10% and 20% sequence identity for amino acids sequences (61).

Global versus local alignments

Needleman and Wunsch (39) published the first global alignment algorithm in 1970. Global alignment algorithms aim at aligning the entire sequence of two potentially homologous regions, as opposed to local alignment algorithms, which align only regions of high similarity. The first local alignment algorithm using dynamic programming was developed by Smith and Waterman (56) as a variation of the Needleman-Wunsch algorithm. The main difference of the Smith-Waterman algorithm is that the alignment can end anywhere in the matrix. By matrix, we refer to the two-dimensional array where we

Fig 1a

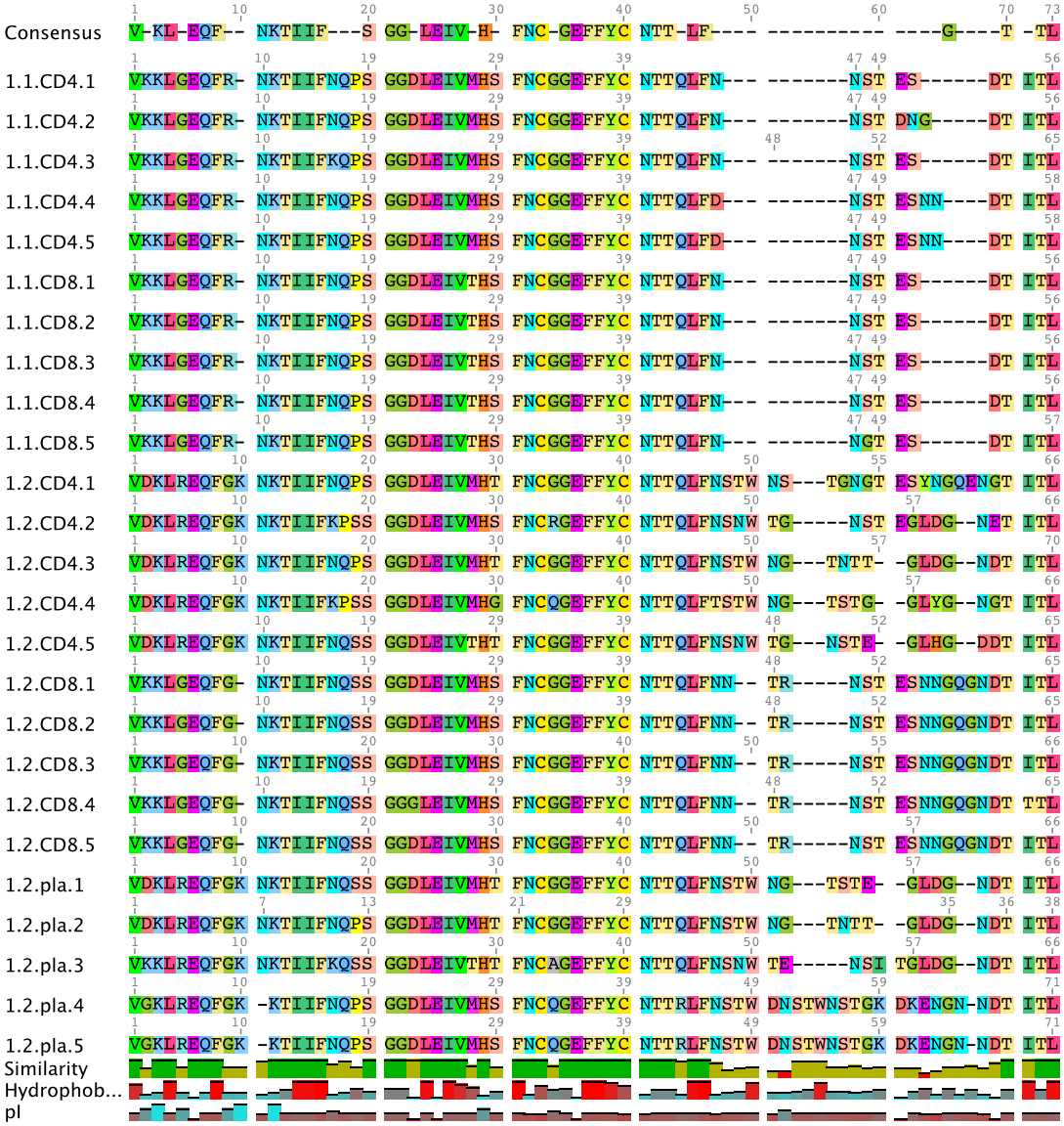
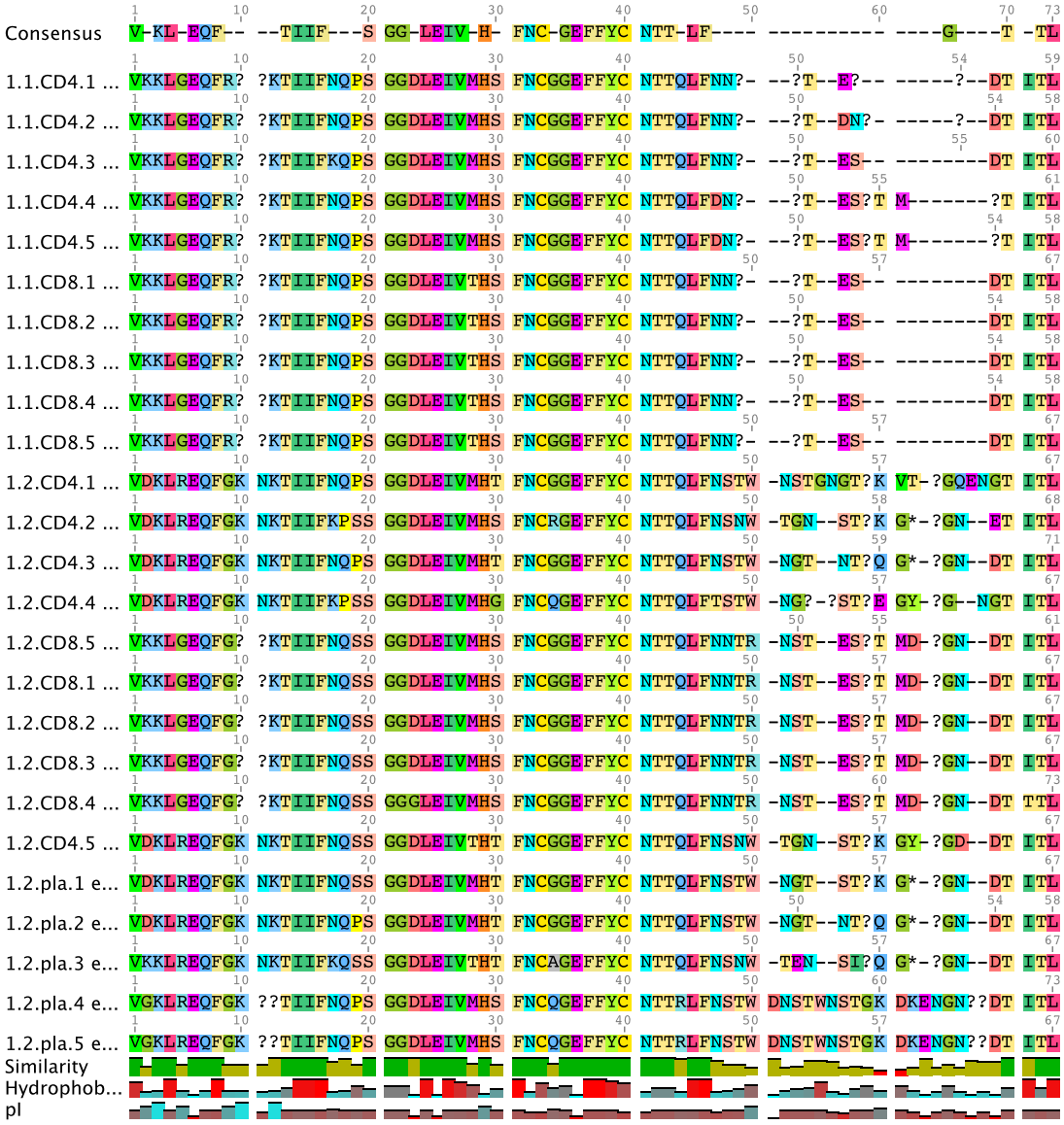


Figure 1 - Amino acid alignment of HIV-1 envelope sequences sampled from a single host. a) ClustalW output (default settings) for the aligned amino acid sequences.

can score identities and differences between two sequences (see Figure 2b). By allowing the alignment to end anywhere in the matrix, only the more similar subsequences of the sequences are aligned. Local alignment is the most sensitive procedure to detect similarity when comparing highly divergent sequences and is therefore very useful for finding common domains between protein sequences or for comparing extended sections of genomic DNA sequences (9). For details on these algorithms, we refer to the original papers (39, 56), books (28, 47), and reviews (49). General descriptions and exercises can also be found in chapter 3 of "The Phylogenetic Handbook" (21), which focuses mainly on global multiple alignment strategies and software.

Fig 1b



b) ClustalW output (default settings) for the aligned nucleotide sequences, translated to amino acid. The pdf images of the alignments were generated using Geneious v2.5.3.

Pairwise alignments

Global pairwise alignment

Global pairwise alignment is, as mentioned above, achieved through the use of the Needleman and Wunsch algorithm (39). This algorithm assumes that the two sequences are similar enough over their entire length to generate a good alignment. In the nucleotide alignment matrix, a positive score is given if there is a match between the sequences and a score of 0 if there is a mismatch. If there is a need to include gaps in the alignment, gap-opening and gap-extending penalties are accounted for in the alignment score. As part of the dynamic programming procedure, the entries of the alignment matrix are computed recursively using the forward algorithm. The back-tracing algorithm is subsequently used to find the best-scoring alignment, starting from the $(n1,n2)^{th}$ position of the matrix and finishing at the $(0,0)^{th}$ position, where n1 and n2 are the lengths of sequences 1 and 2, respectively (see Figure 2) (35, 39, 47).

Local pairwise alignment

It is common practice to compare a query sequence to a database of sequences, in order to find the most similar and potentially homologous sequences. The use of this technique can also assist in quality control of sequencing, by identifying potential lab contaminations, or in compiling an appropriate dataset for further evolutionary studies.

The Smith-Waterman algorithm implements a very straightforward variation of the Needleman-Wunsch algorithm, which is to replace the overall score of the alignment by zero if it takes on negative values for all alternative pathways. This simple approach restricts alignment to regions of reasonably high similarity. How this differs from global alignment is illustrated in the grey box of Figure 2a. Exact alignment algorithms are so computationally expensive that they become unrealistically slow if one wants to compare a sequence to a background database of sequences. To overcome this limitation to database applications, heuristic alignment algorithms have been developed. The most widely used local

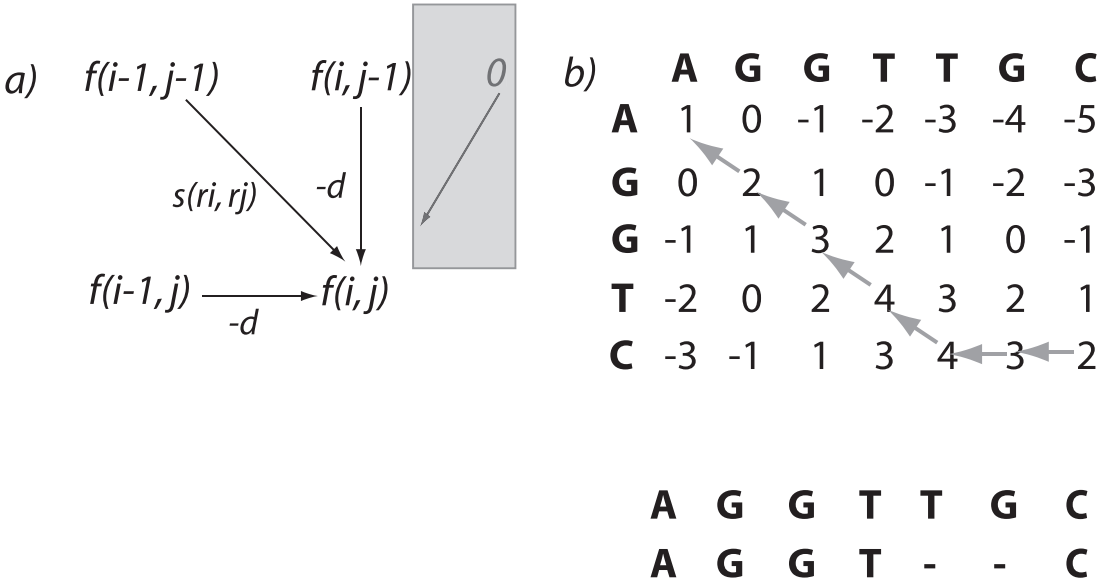


Figure 2 - a) Forward algorithm of the Needleman and Wunsch algorithm to recursively compute the entries of the alignment matrix. The grey box represents the additional parcel of the Smith Waterman algorithm (adapted from (35)) b) Example of an alignment matrix and the back-tracing algorithm used to find the best-scoring alignment.

alignment search tools are BLAST (1) and FASTA (48). The BLAST programs are available at the NCBI website (http://www.ncbi.nlm.nih.gov/BLAST/). The FASTA programs are available at the EMBL-EBI website (http://www.ebi.ac.uk/fasta/). In both cases, the search tool can be used online or downloaded as standalone software to a local computer. The advantage of downloading any of these packages to a local computer is that one can search against a local background database of sequences. This can be very useful if one is interested in searching only against a specific set of sequences. Furthermore, the Los Alamos HIV database (http://www.hiv.lanl.gov/) implements an HIV-BLAST program, which can either use the whole Los Alamos HIV database (default) or a user-defined database as background.

BLAST is usually preferred over FASTA since it is generally faster and more sensitive (because BLAST, in contrast to FASTA, does not require a perfect match in the first step). However, FASTA has been shown to generally perform better than BLAST in terms of mean average precision. For both methods, the search quality appears to be proportional to the logarithm of search time (3).

Multiple sequence alignments

Multiple sequence alignment involves global alignment of more than two sequences. Given that pairwise alignment tries to find the best path in a matrix, multiple sequence alignment can be conceived as a multidimensional problem. A naïve solution to this problem has complexities concerning computation time and memory, which are prohibitively large for real-world alignments (37). The most commonly used heuristic approach to multiple sequence alignment is the 'progressive alignment' algorithm, as originally referred to by Feng and Doolittle (12). Although there are several different progressive alignment strategies (23, 25, 57), the heuristic for aligning the most similar pairs of sequences first is crucial for their performance. Most strategies start from pairwise alignments and use the pairwise similarities to build quick guide trees, thereby reducing the problem of multiple alignment to a set of pairwise alignments (with one alignment at each internal node of the guide tree). This strategy relies on the evolutionary principle that the insertion of gaps should be more straightforward for more closely related sequences. The ideal alignment will be the one that maximizes the sum of similarities for all pairs of sequences. Progressive alignment algorithms generate fast and, in most cases, reasonably accurate results. However, a heuristic algorithm does not guarantee retrieval of the best alignment (42).

Multiple alignment software

We provide a detailed listing of software implementing different alignment algorithms in Table 1, but our discussion below will focus on the most popular or useful programs for HIV according to our experience.

Clustal

The most widely used multiple alignment programs are ClustalW (59) and ClustalX (58). In the Clustal algorithm, sequences are aligned in pairs to generate a distance matrix that can be used to make a rough initial tree of the sequences; this rough tree is used to progressively create the multiple alignment according to the branching order in the guide tree (21, 59). ClustalW is a command-line based software, available for Unix, MacOsX, and Windows operating systems. ClustalX uses the same alignment algorithm and has more or less the same features as ClustalW, but has a graphical interface and is considered to be more user-friendly. The Clustal algorithm is implemented in many alignment editing software packages.

T-Coffee

The T-Coffee multiple alignment software implements an algorithm that combines progressive and consistency-based alignment (44), and is available for online usage (http://www.ch.embnet.org/software/TCoffee.html) or for download (http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html). Because it considers information included in a library of pairwise alignments between the input sequences, the algorithm has been reported to be more accurate than ClustalW (see below). Hence, it is considered to be useful for alignments of more divergent sequences. T-Coffee,

Table 1 Software implementing different alignment algorithms for multiple sequence alignment.
Abbreviations: AA = amino acid, NA = nucleic acid, OS = operating system, WI = Web interface, v = version. All these software packages are available as freeware.

Name	AA/NA Both	Ref.	Algorithm	Website	Supported OS Win	Linux	Mac
ABA	AA	(53)	Progressive/A-Bruijn graphs	http://nbcr.sdsc.edu/euler/aba_v1.0/		X	X
ClustalW	Both	(59)	Progressive	http://bips.u-strasbg.fr/fr/Documentation/ClustalX/	X	X	X
DiAlign	Both	(38)	Consistency-based/Iterative	http://bibiserv.techfak.uni-bielefeld.de/dialign/	X	X	X
MAFFT	Both	(27)	Progressive/iterative	http://align.bmr.kyushu-u.ac.jp/mafft/software/	X	X	X
MSA	Both	(34)	Exact (Carrillo-Lipman optimal alignment algorithm)	http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html		X	
MultAlign (WI)	Both	(5)	Progressive	http://bioinfo.genopole-toulouse.prd.fr/multalin/multalin.html	X	X	X
MUSCLE	Both	(10)	Progressive/Iterative	http://www.drive5.com/muscle/	X	X	X
PileUp (algorithm implemented in SeqLab see Table 2)	Both	(24)	Progressive	http://www.accelrys.com/products/gcg/	-	-	-
POA (WI: POAVIZ or POA v2 for download)	Both	(32)	Progressive (Partially-ordered graphs)	http://www.bioinformatics.ucla.edu/poa		X	
PROBCONS (1.10-1 for download or WI)	AA	(7)	Consistency-based/Probabilistic modeling	http://probcons.stanford.edu/about.html	X	X	X
ProDA (v1.0)	AA	(50)	Progressive/Consistency-based	http://proda.stanford.edu	?	X	X
PRRP (WI)	Both	(16)	Iterative/Stochastic	http://prrn.hgc.jp/	X	X	X
SAGA (v0.95)	AA	(43)	Iterative/Stochastic/GA	http://www.tcoffee.org/Projects_home_page/saga_home_page.html		X	
SAM (v3.5)	AA	(29)	Iterative/Stochastic/HMM	http://rph@cse.ucsc.edu		X	
T-Coffee (v5.03 for download or WI)	Both	(44)	Consistency-based/Progressive	http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html	X	X	X

however, has the disadvantage of being much more time- and memory-consuming. A recent version of this algorithm can also incorporate three-dimensional structural information for the alignment of protein sequences (45).

MAFFT and MUSCLE

The alignment algorithms implemented in MAFFT and MUSCLE are also considered to have good performance characteristics (10, 26). These include variations of the progressive alignment method, which make the multiple alignment process faster than ClustalW and, in some cases, also more accurate (11) (63).

ProbCons

According to Wallace et al., the currently most accurate method for multiple sequence alignment is ProbCons (63). This method uses a pair-hidden Markov Model to specify the probability distribution over all alignments between a pair of sequences. The expected accuracy function is used as a measure of similarity to build the guide tree, and the novel probabilistic consistency scoring function is used for scoring the multiple alignments. However, the reported high accuracy of this method is achieved at the cost of computation time (7).

Comparison of multiple alignment programs

As shown in Table 1, various algorithms for multiple sequence alignment are now available. Independent evaluation of different alignment algorithms has therefore become an important issue. The database mainly used for this type of evaluation is BAliBASE - Benchmark Alignment dataBASE for the evaluation of multiple alignment programs (62). This database was created in 1999 and has since been improved; the current version is BaliBASE v3.0 (60). The alignments included in BAliBASE are divided into four hierarchical reference sets. Each of the main sets may be further subdivided into smaller groups, according to sequence length and percent similarity. The multiple alignments included in BAliBASE were manually refined and therefore constitute an ideal reference alignment set for comparison with the alignments being evaluated (62) (http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/).

To our knowledge, only two papers have been published that independently compare the performance of different algorithms for multiple alignments using BAliBASE. The first paper was published in 1999 and compares the performance of 10 different alignment algorithms: PRRP, ClustalX, SAGA, DiAlign, SB_PIMA, ML_PIMA, MultAlign, PileUp, MULTAL and HMMT.

In the BAliBASE, there are two reference sets that might be interesting for the analysis of the performance of these algorithms for HIV-1 datasets: one of the reverse transcriptase family (1rthA) and another of HIV-1 protease (1fmb). Both are included in reference 1, in the group of alignments with more than 35% identity. The first is included in the subgroup of long sequences and the second in the subgroup of short sequences. In these groups, there were no significant differences reported between the alignment algorithms studied. However, it is shown that the performance clearly decreases with the decrease of percent identity and with shorter sequence lengths. Considering the individual results of the two above-mentioned datasets, ClustalX and PileUp presented a slightly higher score than the other algorithms in the 1rthA dataset, while ClustalX, SAGA, MultAlign and PileUp seem to perform better in the 1fmb dataset (http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/prog_scores.html) (61).

A more recent study by Lassman et al. compared the performance of Poa, DiAlign, T-Coffee and ClustalW. The general conclusion of this paper was that T-Coffee and DiAlign performed better than Poa and ClustalW. However, the differences were only marginal, and ClustalW performed better than DiAlign in the group where 1fmb and 1rthA datasets are included. Poa generally performed worse, but has the advantage of being much faster than the other algorithms (31).

Alignment of protein vs nucleotide sequences

It is well known that at high divergence the ‘signal-noise ratio’ in protein sequences is much better than in nucleotide sequences. Two random nucleotide sequences of equal base composition will be 25%

identical if gaps are not allowed and 50% identical if gaps are allowed (46). This situation may obscure any genuine relationship of homology that may exist at high sequence divergence (46). In addition, amino acid alignments preserve the reading frame of coding sequences, and they also employ more informative scoring matrices, which can increase the quality of the alignment of coding sequences. The example in Figure 1 illustrates the difference between amino acid sequences (Figure 1a) and nucleotide sequences (Figure 1b) as output for a multiple alignment program like ClustalW. In the second case, the information about the reading-frame was lost in the region between 50–70 aa, as seen by the incomplete codons (?) and stop codons (*) present in the alignment.

Despite this advantage, researchers often like to perform further analyses on the protein coding sequences. In this context, the software RevTrans is extremely useful. The program takes as input a set of unaligned nucleotide sequences, translates it, constructs a multiple alignment of the amino acid sequences, and finally builds a multiple alignment of nucleotide sequences by 'reverse translation' of the amino acid alignment (64). This software is available online (http://www.cbs.dtu.dk/services/RevTrans). DAMBE can also be useful for this purpose (see below).

The amino acid scoring matrices have usually been derived from mammalian genome alignments, and their applicability to HIV proteins is not well studied. The development of empirical HIV scoring matrices could therefore be a significant advance in accurately aligning HIV sequences.

Handling gaps

One of the most important problems in editing alignments is handling gaps. Gaps need to be inserted for various reasons, such as indels, sequencing errors, or simply due to different lengths of the sequences. Methods used to infer evolutionary distances can deal with gaps in two different ways: one is to ignore all sites that include gaps or missing data (complete deletion), and the other way is to compute a distance for each pair of sequences, ignoring only gaps involved in the two sequences being compared (pairwise deletion). This second option can be useful if the number of nucleotides involved in the gaps is small and if gaps are distributed randomly in the alignment (40). In likelihood-based phylogenetics, gaps are usually treated as missing data, and maximum likelihood computations average over every possible character state.

The decision on how to treat gaps in an alignment is not always straightforward. Gaps at the beginning or end of the alignment (due to different lengths of the sequences) can be removed by trimming the alignment to the same length. Gap columns in the middle of the alignment can also be removed from the final alignment. However, if a few sequences are much shorter than others or include many gaps, it might not be a good option to include them in the analysis, since a lot of information might be ignored. Therefore, two alternative strategies can be used avoid this: one is to exclude those sequences from the alignment if they are not absolutely necessary for the analysis; the other is to replace each gap in those sequences by the "missing" character (in most software this is represented by a "?") (22).

A good rule of thumb in deciding on what to do with a gap column is to exclude columns where 50% or more of the sequences of the alignment are gapped, while keeping columns where less than 50% of the sequences are gapped.

Alignment editors

Algorithmic alignment does not necessarily retrieve the best alignment. It is important to always verify whether the sequence data are aligned unambiguously and, if necessary, manually correct the alignment. For this purpose, alignment editing software packages are extremely important. A detailed review of all the software for alignment edition, visualization, and presentation is available online at the webpage of the Pasteur Institute (http://bioweb.pasteur.fr/cgi-bin/seqanal/review-edital.pl). We present an extensive listing of available alignment editors in Table 2. We also selected the software that we consider more useful for manual editing of HIV-1 multiple alignments to discuss in more detail below. We selected programs based on their user-friendliness, the availability of additional features, and their availability for Windows, Linux, and MacOsX operating systems. We describe BioEdit, DAMBE, SeAl, GeneDoc, JalView, Geneious, and GDE.

Table 2 Overview of programs for alignment editing. All are freeware with the exceptions of Gene Studio Pro and SeqLab. The academic version of Geneious is free, but doesn't allow alignment editing. The commercial Pro version allows manual alignment editing, ClustalW, and profile alignment. Chromas Lite is a simpler version of Chromas, available free of charge. Chromas can be downloaded for free, but only for a 60-day period.

		Supported operating systems			Input formats														
Editor (version)	Website	Win	Linux	Mac	GenBank	fasta	phylip	nexus	mega	clustal	NBRF pir	GCG msf	RSF	DNAstar	EMBL/Swissprot	GDE	PFAM	DNAstrider	PAML
AnnHyb (4.936)	http://www.bioinformatics.org/annhyb/	X	X			X							X		X				
Base-by-base	http://athena.bioc.uvic.ca/workbench.php?tool=basebybase&db=	X	X	X	X	X				X					X				
BioEdit (7.0.5)	http://www.mbio.ncsu.edu/BioEdit/bioedit.html	X			X	X	X	X		X	X	X		X	X			X	
CINEMA (5)	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.02/index2.html		X								X								
Chromas (2.31)	http://www.infobiogen.fr/services/analyseq	X			X					X		X							
DAMBE (4.5.24)	http://dambe.bio.uottawa.ca/dambe.asp	X			X	X	X	X	X	X	X	X			X			X	X
DNA Stacks (1.3.4)	http://biology.fullerton.edu/deernisse/dnastacks.html			X	X	X	X	X		X			X		X				
GDE (2.2)	http://www.bioafrica.net/GDElinux/index.html		X	X	X	X	X	X		X	X	X	X			X		X	
Se-Al (2.0a11)	http://evolve.zoo.ox.ac.uk/software.html?name=Se-Al			X		X	X	X			X					X			
GeneDoc (2.6.03)	http://www.psc.edu/biomed/genedoc/	X				X				X	X	X							
Geneious (2.5.3)	http://www.geneious.com/	X	X	X	X	X		X	X	X		X	X	X				X	
GeneStudio Pro (2.0.6.3)	http://www.genestudio.com/genestudio.htm	X			X	X	X	X		X	X	X		X	X	X		X	
JalView (2.2)	http://www.jalview.org/	X	X	X		X				X	X	X					X		
JaMBW (1.1)	http://www.infobiogen.fr/services/analyseq	X	X	X	X	X					X				X				
JevTrace (3.12b)	http://www.cmpharm.ucsf.edu/%7Emarcinj/JEvTrace/index.html	X	X	X		X				X			X						
MUST 2000	http://www.isem.univ-montp2.fr/PPP/PM/RES/Info/@Softwares.php#MUST2000	X			X						X				X				
ProMSED (2)	ftp://ftp.ebi.ac.uk/pub/software/dos/promsed/	X				X				X	X	X	X		X		X		
SeaView	http://pbil.univ-lyon1.fr/software/seaview.html	X	X	X		X	X	X		X		X							
GCG/SeqLab	http://www.accelrys.com/products/gcg/	X	X	X	X	X					X				X				
SeqPup (0.9)	http://iubio.bio.indiana.edu/soft/molbio/seqpup/java/seqpup-doc.html	X	X	X	X	X	X	X		X	X	X			X			X	
W2H (4.1.2) (web interface)	http://www.w2h.dkfz-heidelberg.de/	X	X	X	X	X					X				X				

BioEdit

BioEdit is a manual alignment editor available for Windows. It includes many analysis tools for sequences/alignments and allows several external tools to be configured to run through the BioEdit interface. Examples of the external tools that can be run through the BioEdit interface are TreeView, BLAST, and ClustalW. Many other useful features are incorporated into BioEdit, such as the ability to obtain consensus sequences, amino acid and nucleotide composition statistics, entropy plots, hydrophobicity profiles, and dot plots of pairs of sequences. The last update of BioEdit was in May 2005 (v7.0.5), but it is still available online for download (see Table 2 for link)(19).

DAMBE

DAMBE (Data Analysis in Molecular Biology and Evolution) is an integrated Windows program for descriptive and comparative analysis of molecular data. It has features for manipulating, editing, and converting sequences and alignments in different formats. One of the most interesting features of DAMBE is its ability to align protein-coding nucleotide sequences against aligned amino acid sequences, which avoids frame-shifts typical for alignments of HIV nucleotide sequences (see Figure 1b). In addition, DAMBE can compute statistics for nucleotide, amino acid, and codon frequencies, as well as codon usage and amino acid usage bias. Moreover, the program implements several comparative sequence analysis features: quantification of substitution patterns and fitting statistical distributions to among-site substitution rate heterogeneity, therefore helping in the selection of a substitution model; phylogenetic reconstruction based on distance, maximum-parsimony and maximum-likelihood methods with the option to perform bootstrapping and jackknifing; testing alternative phylogenetic hypotheses; phylogenetic tree viewing and manipulation; and graphical tools as well as formal tests to evaluate the phylogenetic signal of a dataset (65).

Se-Al

Se-Al is a very straightforward and user-friendly software strictly focused on alignment editing, visualization, and file format conversion. It allows easy toggling between nucleotide reading frames, codons, and amino acids, and presents sequences with appropriate coloring, allowing interactive editing of the sequence alignment. Thanks to editing features like selecting and sliding individual residues or blocks of sequence stretches, and cutting and pasting, which can be performed with the usual MacOS short keys, it is an easy and flexible tool for manually editing alignments. Other useful Se-Al features are the ability to generate consensus sequences, switching between reverse, complement and reverse-complement of selected sequences, selecting site ranges, and several gap deletion options (52).

GeneDoc

GeneDoc provides tools for visualizing, editing, and analyzing multiple sequence alignments. The program can incorporate structural or biochemical information as guidance to which residues should be aligned, and secondary structures can be easily visualized. GeneDoc can perform pairwise alignment and multiple sequence alignment, and allows the user to compute the score of the alignment for any selected fragment. GeneDoc also provides some additional analysis tools, which can be useful for grouping sequences in the alignment. These include the Kolmogorov-Smirnov tests of distributions of alignment scores or comparisons of sequences in terms of the percentage of identities between a pair of aligned sequences. A positive result in the test of whether the scores for pairs of sequences within the same group are smaller than the scores for pairs of sequences that are in different groups would indicate that the grouping categories are systematically reflected in the sequences (41).

JalView

In addition to the general viewing and editing features of JalView, the program has several useful features for the analysis of sequences or alignments. JalView is one of the few programs that can realign sequences using three different alignment algorithms: MUSCLE, MAFFT, or ClustalW. JalView can also make pairwise alignments of user-selected sequences, build UPGMA and NJ trees based on percent identity distances, cluster sequences based on principal component analysis (PCA), and perform secondary structure prediction (4).

Geneious

Geneious is a software package recently developed by the bioinformatics company Biomatters. The fully featured pro version is commercially available, but a limited academic version is freely available for download (http://www.geneious.com). The free version of Geneious offers some visualization and analysis tools for sequences or alignments, including an interface where one can integrate sequences, alignments, 3D structures, and tree data. The editing tools, however, are only available in the pro version. The software connects to public databases for retrieval of datasets and performance of BLAST searches. Geneious is therefore an excellent tool to manage large amounts of data.

Additional analysis tools have been implemented: multiple and pairwise sequence alignment, phylogenetic tree building, dot plots, consensus sequences, and statistics of residue frequencies, pairwise similarity, etc. Furthermore, Geneious allows the user to install custom-designed plug-ins, like a PhyML plug-in for fast maximum likelihood phylogenetic reconstruction (18), a MrBayes plug-in for Bayesian phylogenetic inference (55), and a plug-in for calculating Shannon Entropy Scores (8).

GDE

GDE (Genetic Data Environment) is a user-friendly interface that integrates different bioinformatics tools, without the need to convert between different input/output file formats every time a new tool is used. GDE is very useful for the analysis of HIV, as all of the sequence-specific databases, phylogenetic datasets, and programs needed to study its diversity and molecular phylogeny can be integrated into this interface. Some of the external tools included in GDE are BLAST and FASTA for local alignment against databases, Clustal for MSA, ReadSeq for format conversion, and Phylip for phylogenetic analyses (6).

Alignment tools specific for HIV and HIV pre-built alignments

The Los Alamos HIV Sequence Database website provides several useful web tools for the manipulation of HIV sequences. Sequence Locator can be used to find the location of HIV or SIV nucleotide or protein sequences according to the standard numbering of the HXB2 genome. Gene Cutter extracts HIV genes from nucleotide sequences, which can be pre-aligned. Gene Cutter can codon-align HIV sequences, as well as translate them to amino acids. SynchAlign is another very useful tool to profile-align two alignments, which can include the pre-built HIV alignments available from this database. Primalign and Epilign align nucleotide or protein sequences, respectively, to these same pre-built HIV alignments.

The Los Alamos HIV Database also provides many other tools for working with sequences, which are not exclusive to HIV, but can also be used for other organisms. These include Gapstreeze for removing alignment columns if they present gaps above a certain percentage, Translate for translating nucleotide sequences to amino acids and Format Converter, which accepts sequences in any format and converts them into any other format. More information can be found at the Los Alamos database website.

As mentioned before, the Los Alamos HIV Database provides pre-built alignments of HIV and SIV, of the complete genome, and of specific genomic regions (e.g. LTR, *pol*, *env*, etc) (http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html) (30).

PFAM also provides alignments of retroviral protein families, such as retroviral aspartyl protease (RVP - accession PF00077), reverse transcriptase (RVT - accession PF00078), the retroviral matrix proteins (clan Matrix), and the envelope proteins (GP120 - accession PF00516 and GP41 - accession PF00517). More information can be found at the PFAM website (http://www.sanger.ac.uk/Software/Pfam/) (13).

Publishing alignments

The process of publishing alignments also implies the use of specific software. Not all alignment editors discussed above include functions for outputting ‘pretty-view’ alignments in publishable file formats.

There are also some software packages that were developed specifically for viewing and publishing alignments. These include:

- SeqPublish (http://www.hiv.lanl.gov/content/hiv-db/SeqPublish/seqpublish.html) available online;
- Highlighter (http://www.hiv.lanl.gov/content/hiv-db/HIGHLIGHT/highlighter.html) available online;
- Alscript (http://www.compbio.dundee.ac.uk/Software/Alscript/alscript.html), available for Linux and Windows;
- BOXSHADE (http://www.ch.embnet.org/software/BOX_form.html) available online or for Mac, Windows, and Linux;
- ESPript (http://espript.ibcp.fr/ESPript/ESPript/) available online and for Linux (17);
- STRAP (http://www.charite.de/bioinf/strap/) available for Linux, MacOsX, and Windows (15);

Discussion

Sequence alignment is a necessary prerequisite for the analysis of gene and protein sequence data. These analyses can include phylogenetic inference, structural analysis, and data mining. If a sequence alignment contains errors, these errors will be propagated in subsequent analysis, with the potential to result in flawed conclusions. While local alignment methods are important for searching closely related sequences and for quality control of laboratory-generated sequences, application of global alignment methods is required before proceeding with comparative sequence analysis. Recent developments aim at alignment-free phylogenetic inference (33, 51) or may improve co-estimation of alignment, phylogeny and derived parameters. An interesting development in recent years has been "statistical alignment". This class includes multiple alignment algorithms that use a statistical method, such as hidden-Markov models implemented in a Bayesian approach (2, 36, 54) or other statistically-based methods that attempt to associate a P-value to the multiple alignment (20, 42).

Visually inspecting sequence alignments is currently required to ensure their quality before proceeding to further analyses. Especially in regions with a lot of indels, such as the envelope of HIV, alignments need to be manually edited to improve their quality. In our experience, however, manual editing usually comes down to deleting ambiguously aligned gene regions. For example, hypervariable loops in HIV envelope genes, like the one shown in Figure 1, may need to be omitted from further analysis. A simple guideline would be to delete an ambiguous alignment part in between two conserved residues, for example the hypervariable V4 loop between residues 46 (F) and 70 (T) in Figure 1a. We would like to note that 'gap-stripping', which involves the removal of all alignment columns that contain gaps, and which is a frequent practice in the analysis of viral sequences, should not be considered as a standard prerequisite for further analysis. Regions where gaps have been inserted with relatively high confidence can still be informative in further analysis. For example, in our opinion, there is no need to delete the gapped region from residue 8 to residue 11 in Figure 1. More editing may be needed for nucleotide sequence alignments, in particular to restore the reading-frame in coding sequences (Figure 1b). Many software programs are available for this purpose. We briefly discussed seven of those that are considered to be relatively user-friendly and that cover common operating systems. Unfortunately, there exists no comprehensive software that combines all the useful features of the different programs we discussed. As academic funding agencies generally under-appreciate software development, researchers may sometimes need to resort to commercial software packages that can fill the gap.

Acknowledgments

ABA was supported by Fundação para a Ciência e Tecnologia (Grant nr SFRH/BD/19334/2004). PL was supported by an EMBO long-term fellowship.

References

1. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403-10.

2. Baldi, P., Y. Chauvin, T. Hunkapiller, and M. A. McClure. 1994. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci U S A* **91**:1059-63.
3. Chen, Z. 2003. Assessing sequence comparison methods with the average precision criterion. *Bioinformatics* **19**:2456-60.
4. Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton. 2004. The Jalview Java alignment editor. *Bioinformatics* **20**:426-7.
5. Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**:10881-90.
6. De Oliveira, T., R. Miller, M. Tarin, and S. Cassol. 2003. An integrated genetic data environment (GDE)-based LINUX interface for analysis of HIV-1 and other microbial sequences. *Bioinformatics* **19**:153-4.
7. Do, C. B., M. S. Mahabhashyam, M. Brudno, and S. Batzoglou. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* **15**:330-40.
8. Drummond, A., M. Kearse, J. Heled, R. Moir, T. Thierer, B. Ashton, A. Wilson, and S. Stones-Havas. 2006. Geneious v2.5. Available from http://www.geneious.com/.
9. Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis - Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK.
10. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792-7.
11. Edgar, R. C., and S. Batzoglou. 2006. Multiple sequence alignment. *Curr Opin Struct Biol* **16**:368-73.
12. Feng, D. F., and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25**:351-60.
13. Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. 2006. Pfam: clans, web tools and services. *Nucleic Acids Res* **34**:D247-51.
14. Gardner, P. P., A. Wilm, and S. Washietl. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* **33**:2433-9.
15. Gille, C., and C. Frommel. 2001. STRAP: editor for STRuctural Alignments of Proteins. *Bioinformatics* **17**:377-8.
16. Gotoh, O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol* **264**:823-38.
17. Gouet, P., E. Courcelle, D. I. Stuart, and F. Metoz. 1999. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* **15**:305-8.
18. Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**:696-704.
19. Hall, T. 2005. BioEdit version 7.0.5. Available from http://www.mbio.ncsu.edu/BioEdit/bioedit.html. [Online.]
20. Hertz, G. Z., and G. D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**:563-77.
21. Higgins, D. G. 2003. Multiple alignment: Theory. In M. Salemi and A.-M. Vandamme (ed.), The Phylogenetic Handbook - A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge.
22. Higgins, D. G., and M. Salemi. 2003. Multiple alignment: Practice. In M. Salemi and A.-M. Vandamme (ed.), The Phylogenetic Handbook - A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge.
23. Higgins, D. G., and P. M. Sharp. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**:237-44.
24. Higgins, D. G., and P. M. Sharp. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosci* **5**:151-3.
25. Hogeweg, P., and B. Hesper. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J Mol Evol* **20**:175-86.

26. Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**:511-8.
27. Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**:3059-66.
28. Korf, I., M. Yandell, and J. Bedell. 2003. BLAST - An Essential guide to the basic local alignment search tool, 1st ed. O'Reilly & Associates, Sebastopol, CA, USA.
29. Krogh, A., M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**:1501-31.
30. Los Alamos National Laboratory. Los Alamos Database. http://www.hiv.lanl.gov/content/index. [Online.]
31. Lassmann, T., and E. L. Sonnhammer. 2002. Quality assessment of multiple alignment programs. *FEBS Lett* **529**:126-30.
32. Lee, C., C. Grasso, and M. F. Sharlow. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**:452-64.
33. Li, M., J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* **17**:149-54.
34. Lipman, D. J., S. F. Altschul, and J. D. Kececioglu. 1989. A tool for multiple sequence alignment. *Proc Natl Acad Sci U S A* **86**:4412-5.
35. Liu, J., and T. Longvinenko. 2003. Bayesian methods in biological sequence analysis, Handbook of Statistical Genetics, 2nd ed, vol. 1. John Wiley & Sons, Ltd., West Sussex.
36. Lunter, G., I. Miklos, A. Drummond, J. L. Jensen, and J. Hein. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6**:83.
37. McNaughton, M., P. Lu, J. Schaeffer, and D. Szafron. 2002. Presented at the 18th National Conference on Artificial Intelligence.
38. Morgenstern, B., K. Frech, A. Dress, and T. Werner. 1998. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14**:290-4.
39. Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**:443-53.
40. Nei, M., and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.
41. Nicholas, K., H. Nicholas, and D. Deerfield. 1997. GeneDoc: Analysis and Visualization of Genetic Variation. Available from http://www.psc.edu/biomed/genedoc. [Online.]
42. Notredame, C. 2002. Recent progress in multiple sequence alignment: a survey. Pharmacogenomics **3**:131-44.
43. Notredame, C., and D. G. Higgins. 1996. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res* 24:1515-24.
44. Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**:205-17.
45. O'Sullivan, O., K. Suhre, C. Abergel, D. G. Higgins, and C. Notredame. 2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* **340**:385-95.
46. Opperdoes, F. 2003. Phylogenetic analysis using protein sequences: Theory. In M. Salemi and A.-M. Vandamme (ed.), The Phylogenetic Handbook - A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge.
47. Pearson, W., and T. Wood. 2003. Statistical Significance in Biological Sequence Comparison, Handbook of Statistical Genetics, 2nd ed, vol. 1. John Wiley & Sons, Ltd., West Sussex.
48. Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**:2444-8.
49. Phillips, A. J. 2006. Homology assessment and molecular sequence alignment. *J Biomed Inform* **39**:18-33.
50. Phuong, T. M., C. B. Do, R. C. Edgar, and S. Batzoglou. 2006. Multiple alignment of protein sequences with repeats and rearrangements. *Nucleic Acids Res* **34**:5932-42.

51. Potter SJ, Lemey P, Dyer WB, Sullivan JS, Chew CB, Vandamme AM, Dwyer DE, Saksena NK. Genetic analyses reveal structured HIV-1 populations in serially sampled T lymphocytes of patients receiving HAART. *Virology*, **348**(1):35-46.
51. Qi, J., B. Wang, and B. I. Hao. 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J Mol Evol* **58**:1-11.
52. Rambaut, A. 1996. Se-Al: Sequence Alignment Editor. Available at http://evolve.zoo.ox.ac.uk/
53. Raphael, B., D. Zhi, H. Tang, and P. Pevzner. 2004. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Res* **14**:2336-46.
54. Redelings, B. D., and M. A. Suchard. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst Biol* **54**:401-18.
55. Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572-4.
56. Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* **147**:195-7.
57. Taylor, W. R. 1988. A flexible method to align large numbers of biological sequences. *J Mol Evol* **28**:161-9.
58. Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**:4876-82.
59. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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**:4673-80.
60. Thompson, J. D., P. Koehl, R. Ripp, and O. Poch. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* **61**:127-36.
61. Thompson, J. D., F. Plewniak, and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res* **27**:2682-2690.
62. Thompson, J. D., F. Plewniak, and O. Poch. 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**:87-8.
63. Wallace, I. M., G. Blackshields, and D. G. Higgins. 2005. Multiple sequence alignments. *Curr Opin Struct Biol* **15**:261-6.
64. Wernersson, R., and A. G. Pedersen. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res* **31**:3537-9.
65. Xia, X., and Z. Xie. 2001. DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* **92**:371-3.