Visualization of multiple alignments, phylogenies and gene family evolution

James B Procter[1], Julie Thompson[2], Ivica Letunic[3], Chris Creevey[4], Fabrice Jossinet[5] & Geoffrey J Barton[1]

Software for visualizing sequence alignments and trees are essential tools for life scientists. In this review, we describe the major features and capabilities of a selection of stand-alone and web-based applications useful when investigating the function and evolution of a gene family. These range from simple viewers, to systems that provide sophisticated editing and analysis functions. We conclude with a discussion of the challenges that these tools now face due to the flood of next generation sequence data and the increasingly complex network of bioinformatics information sources.

Tree and sequence alignment visualizations have a long history. Evolutionary tree diagrams can be found in even the earliest descriptions of evolution, and their visualization still plays a key role in modern phylogenetics. However, although trees visualize an organism's evolutionary history, it is the biological data used in their construction that contains the information that distinguishes each organism. Sequence alignments are the most common data used in phylogenetic analysis, and their visualization assists in understanding the molecular mechanisms that differentiate each species, down to the level of the individual nucleotide bases and amino acids.

Many tools for tree and sequence alignment visualization have been developed in the last 20 years, and a comprehensive analysis is beyond the scope of this review. Instead, we describe the main visualization approaches found in a selection of applications that are available at present (**Tables 1** and **2**), and that we consider either to be widely used or to represent a significant contribution to each field. We also highlight important capabilities and drawbacks for each tool, but since many are under active development, we urge the user to explore a tool's capabilities for themselves.

Several functions can be found among the tree and alignment visualization tools we consider here: 'renderers' generate static figures, 'viewers' allow interactive display and analysis, and 'workbenches' provide a complete environment for creation, visualization, editing, annotation and analysis. Some tools are more specialized and provide functions essential for editing and analyzing alignments or working with RNA (**Table 1**) or allow the user to map other kinds of biological data ('Annotators', **Table 2**).

SEQUENCE DATABASE SEARCHES

Many sequence analysis exercises begin by using a search tool such as BLAST[1]. These tools use fast alignment methods to compare a query sequence against a library of potential sequence or sequence-family matches. The result is a ranked list of query–hit alignments, each with an associated alignment score and estimate of the significance of the match. The user is then tasked with examining this list, to identify the alignments relevant to their investigation for use in the next stage of the analysis.

Probably the most widely used visualization tool for sequence database search results is the BLAST viewer[2] at the US National Center for Biotechnology Information (NCBI) website. This web-based system has its roots in the textual report generated by BLAST search tools. However, the main advantage of this viewer is that it provides a summary diagram that gives a bird's-eye view of the aligned positions of each hit on the query sequence. Each hit is colored by the bit score for its match to the query to indicate alignment quality, and a hyperlink takes the viewer to the pairwise alignment, enabling

[1]School of Life Sciences Research, College of Life Sciences, University of Dundee, Dundee, UK. [2]Institute of Genetics and Molecular and Cellular Biology (IGBMC), Strasbourg, France. [3]European Molecular Biology Laboratory, Heidelberg, Germany. [4]Animal Bioscience Centre, Teagasc, Ireland. [5]Architecture et réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du Centre National de la Recherche Scientifique (CNRS), Strasbourg, France. Correspondence should be addressed to J.B.P. (j.procter@dundee.ac.uk).

PUBLISHED ONLINE 1 MARCH 2010; DOI:10.1038/NMETH.1434

Table 1 | Selected tools for multiple sequence alignment visualization

Name	Cost[a]	OS[c]	Function[d]	Description	URL
Stand-alone					
AlScript[9]	Free	Win, Mac, Linux	Renderer	Powerful layout engine but complex control files	http://tinyurl.com/pol2ta/
BOXShade	Free	Win, Mac, Linux	Renderer	Simple web interface, basic alignment figures	http://tinyurl.com/mxf6nd/
ClustalX2[4]	Free	Win, Mac, Linux	Viewer	User interface to ClustalW	http://www.clustal.org/
VISSA[37]	Free	Win, Mac, Linux	Viewer	Mapping between alignments and protein structures	http://tinyurl.com/quxjt7/
BioEdit	Free	Win	Edit & anal.	Nucleic acid sequence alignment tools	http://tinyurl.com/nofcdr/
Cinema[24]	Free	Win, Mac, Linux	Edit & anal.	Motif autogeneration; part of Utopia[50]	http://tinyurl.com/rxjb8e/
GeneDoc	Free	Win	Edit & anal.	Visualizer for MEME[21] motif discovery results	http://tinyurl.com/om6mfm/
Jalview[25]*	Free	Win, Mac, Linux	Edit & anal.	Interactive annotation; linked tree views	http://www.jalview.org/
Jevtrace[35]	Free	Win, Mac, Linux	Edit & anal.	Automated tree subfamily analysis	http://tinyurl.com/n6jvr5/
PFAAT[18]*	Free	Win, Mac, Linux	Edit & anal.	Quantitative annotation modeling	http://pfaat.sourceforge.net/
SeaView[23]	Free	Win, Mac, Linux	Edit & anal.	Lightweight, 'guided' alignment editing	http://tinyurl.com/oddnsa/
4SALE[59]	Free	Win, Mac, Linux	RNA	RNA structure prediction tools	http://tinyurl.com/o2n97e/
ConStruct[60]	Free	Mac, Linux	RNA	Handles pseudoknots	http://tinyurl.com/lo63cd/
S2S[61]	Free	Win, Mac, Linux	RNA	Requires RNA reference structure	http://tinyurl.com/qeda97/
SARSE[62]	Free	Mac, Linux	RNA	Semiautomated alignment refinement	http://sarse.kvl.dk/
eBioX[11]	Free	Mac	Workbench	Access to extensive tool set including EMBOSS suite	http://tinyurl.com/yldee4e/
Geneious*	$	Win, Mac, Linux	Workbench	Innovative BLAST result viewer	http://www.geneious.com/
MacVector[19]	$	Mac	Workbench	Spreadsheet for recording analysis results	http://www.macvector.com/
SeqPad	Free	Win, Mac, Linux	Workbench	Built on biojava platform; not yet stable	http://trac.seqpad.org/
Strap[63]	Free	Win, Mac, Linux	Workbench	Sequence and structure alignment and analysis	http://3d-alignment.eu/
Vector NTI[3]	$[b]	Win, Mac, Linux	Workbench	Sophisticated annotation diagrams	http://tinyurl.com/c92xm7/
Web-based					
Chroma+[10]	Free		Renderer	Supports JOY[64] annotated alignments	http://www.llew.org.uk/chroma/
ESPript[8]	Free		Renderer	Rendering engine for ENDSCRIPT	http://tinyurl.com/m5dqwd/

[a]Free means the tool is free for academic use; $ means the tool is not free, but a demo version is available. [b]Demo version is severely limited. [c]OS, operating system: Win, Microsoft Windows; Mac, Macintosh OS X. Tools running on Linux usually also run on other versions of Unix. [d]Renderers are tools that generate figures via the web or command line only. Viewers are interactive alignment visualization tools without editing capabilities. Edit & anal.: editing, annotation and analysis tools. RNA, tools with special support for RNA alignment editing and analysis. Workbench tools are often aimed at experts and provide a range of analysis and visualization functions in addition to alignment visualization. *Our recommendations.

deeper inspection. The NCBI interface has been enhanced over the years, to keep pace with the increasing size of sequence databases, and now provides a tree representation of the search results, so that the relationships within the hit set may be seen. Furthermore, the annotation of each hit with an NCBI taxonomy identifier allows a phylogenetic breakdown of the hit list to be displayed, so the researcher can focus on the query's similarity to sequences found in a single organism, or specific clades.

There are surprisingly few alternatives to the style of visualization provided by the NCBI viewer. However, two of the alignment workbench tools in **Table 1** include innovative approaches (**Supplementary Fig. 1**). Vector NTI[3] presents a fivefold linked view: a hierarchical list giving the details of each hit, the hit summary diagram, a pairwise alignment view, the currently selected hit's alignment trace showing the corresponding homologous segments in both sequences, and a two-dimensional plot of the hit profile on the query sequence. Hit selection is facilitated by the plot, and sequence region selection is possible in either alignment view. Geneious's BLAST viewer provides a 'Linnaeus view' in addition to a more conventional multiple alignment view (see below). In the Linnaeus view, hits are shown as a two-dimensional taxonomic tree map with the 'top hit' identified within its clade by an arrow. Each cell contains a distance tree for hits in that clade, with cells grayed for clades with hits below a user-defined threshold.

MULTIPLE SEQUENCE ALIGNMENTS

A multiple sequence alignment (MSA) is a matrix, in which each row corresponds to a sequence and each column defines equivalent positions across all sequences. Sequence search results are a collection of alignments, and they must usually be transformed into a single MSA before further analysis. BLAST, and some of the workbenches in **Table 1**, can create an MSA by aligning each hit to the query using the pairwise alignment from the search results (**Fig. 1a**), but this approach often introduces errors. True multiple-alignment algorithms obtain more accurate information with a variety of optimization heuristics, such as the guide tree approach (**Fig. 1b**) found in ClustalW[4] and the consistency method (**Fig. 1c**) used by T-Coffee[5], and these are extensively reviewed elsewhere[6,7].

Multiple alignment renderers

Many tools for multiple alignment visualization have been developed over the years. Each one offers some variant of the spreadsheet-style alignment diagram shown in **Figure 2a**. Here, sequences are laid out in rows, and corresponding residues and bases are represented as letters arranged on a grid. Renderers (such as ESPript[8], ALSCRIPT[9] and Chroma[10]) were the first dedicated systems to generate these kind of visualizations, and although appearing outdated by twenty-first-century standards, they still provide the greatest control for automated figure creation. In addition to parsing sequence alignment files, they take a set of parameters via either the web interface,

Table 2 | Selected tools for phylogenetic tree visualization[a]

Name	OS[b]	Function[c]	Description	URL
Stand-alone				
TreeDyn[46]*	Win, Mac, Linux	Renderer	Turnkey tree editor and annotator	http://www.treedyn.org/
Archaeopteryx[65]*	Win, Mac, Linux	Viewer	Viewer/editor providing reference support for phyloXML[d]	http://tinyurl.com/c9vp2d/
CTree[66]	Win, Mac, Linux	Viewer	Viewer for analysis and visualization of clusters within trees	http://tinyurl.com/pd3m3l/
Dendroscope[67]*	Win, Mac, Linux	Viewer	Interactive viewer for large phylogenetic trees and networks	http://tinyurl.com/2etsd8/
FigTree	Win, Mac, Linux	Viewer	Modern tree viewer with coloring and collapsing	http://tinyurl.com/cjrxcd/
HyperTree[41]	Win, Mac, Linux	Viewer	Simple hyperbolic viewer	http://tinyurl.com/55moet/
NJplot[68]	Win, Mac, Linux	Viewer	Interactive tree plotter; reroots, exports as PDF	http://tinyurl.com/lbjw4x/
Tree Set Viz[69]	Win, Mac, Linux	Viewer	Viewer that computes and visualizes distances between trees	http://tinyurl.com/otvc7g/
TreeView[70]	Win, Mac, Linux	Viewer	Classic tree viewing software that is very highly cited	http://tinyurl.com/nn95wv/
TreeJuxtaposer[71]	Win, Mac, Linux	Viewer	The first viewer implementing the focus+context navigation technique	http://olduvai.sourceforge.net/
Walrus[42]	Win, Mac, Linux	Viewer	Generic 3D hyperbolic viewer; no support for standard phylogenetic formats	http://tinyurl.com/ac4cs/
NOTUNG[39]	Win, Mac, Linux	Annotator	ATV-based tool for ortholog and paralog identification by tree reconciliation[e]	http://tinyurl.com/yhyztd7/
Treebolic	Win, Mac, Linux	Annotator	Generic hyperbolic viewer/editor; no support for phylogenetic formats	http://treebolic.sourceforge.net/
TreeGraph[49]	Win, Mac, Linux	Annotator	Annotate with multiple support values or through different widths and colors	http://treegraph.bioinfweb.info/
Treevolution[47]	Win, Mac, Linux	Annotator	'Distortable' tree layout, subfamily highlighting	http://tinyurl.com/kq22s9/
ARB[72]*	Mac, Linux	Workbench	Complete analysis environment	http://www.arb-home.de/
MEGA[72]*	Win, Mac, Linux	Workbench	Workbench for molecular evolutionary genetics analysis	http://www.megasoftware.net/
Mesquite	Win, Mac, Linux	Workbench	Modular system for evolutionary analysis	http://mesquiteproject.org/
SplitsTree4[73]	Win, Mac, Linux	Workbench	Tree and network creator and viewer	http://tinyurl.com/mpbhsg/
TOPALi[74]	Win, Mac, Linux	Workbench	Nucleic acid and protein evolutionary analysis	http://www.topali.org/
Web-based				
Phylodendron		Renderer	Supports a range of tree and branch styles and output formats	http://tinyurl.com/m3cdqb/
Hypergeny		Viewer	Hyperbolic tree browser	http://tinyurl.com/nhrfbq/
iTOL[48]*		Annotator	Powerful tree-based annotation visualizer; batch interface	http://itol.embl.de/
PhyloWidget[75]		Annotator	Processing-based editor/publisher; annotate with image and web links	http://www.phylowidget.org/

[a]All tools in this table are free for academic use. [b]OS, operating system: Win, Microsoft Windows; Mac, Macintosh OS X. Tools running on Linux usually also run on other versions of Unix. [c]Renderers are tools that generate figures by means of a web or command line interface only. Viewers are tools for interactive visualization that have no tree-generation capabilities. Annotators are viewers that allow additional data to be mapped onto the phylogenetic visualization. Workbench tools can generate, manipulate, analyze and visualize trees. [d]phyloXML is a new format for the exchange of phylogenetic trees. [e]ATV refers to 'Another Tree Viewer'—which has been superseded by Archaeopteryx. 3D, three-dimensional. *Our recommendations.

command line arguments or a separate file. These parameters control how the alignment is drawn and annotated, and some can be defined independently of the MSAs being rendered, facilitating the use of these noninteractive tools in sequence analysis pipelines.

Interactive alignment viewers

Interactive viewers generally provide the same visualizations as static renders but adapted for display on screen. Historically, they were developed as a user-friendly interface to alignment programs (ClustalX[4]), and more recently, for sequence analysis suites (eBioX[11]). Importantly, their interface allows shading and display styles to be easily changed, facilitating figure generation. A further advantage is gained when working with large alignments. For example, an MSA taken from the Pfam[12] protein domain family database contains, on average, 300 sequences[12]—which is too large to interpret without the ability to scroll around the alignment matrix and zoom in or out to perceive and focus in on gross trends.

Coloring. The primary role of color in alignment visualizations is the identification of regions where specific properties predominate and to highlight variation. The simplest way this is achieved is to color each sequence symbol according to a specific amino acid or nucleotide color scheme (**Fig. 2b**,**c**). Schemes are usually one of two types: quantitative schemes convey trends in specific empirical properties, such as hydrophobicity or 'burial' in proteins, and qualitative colorings reflect general physicochemical class (for example, sugar ring geometry or amino acid side chain size, shape, polarity and aromaticity). The assignment of colors according to chemical nature is analogous to the conventions for atom colors prevalent in molecular graphics, and the amino acid colors used by Clustal[4] in alignments (**Fig. 2a**) broadly correspond with the main groupings of physicochemical attributes of the 20 amino acids (**Fig. 2d**). Tools vary in the precise choice of scales and color gradations used for quantitative schemes, but among the qualitative schemes, Taylor[13] and Clustal[4] (**Fig. 2c**) are widely supported and may be considered *de facto* standards.

Shading. Coloring every symbol in an alignment can help identify gross trends, but becomes confusing for regions showing complex patterns of variation. A more effective approach, pioneered in ClustalX[4], is to shade symbols on the basis of both their type and their predominance at each alignment position (**Fig. 2a**). This approach is widely supported; and it has many variations, as other measures can be used to define color or control shading, such as a symbol's similarity to some reference (usually the consensus or the sequence used for a BLAST search). Alignment quality can also be emphasized. Dissimilar sites can be rendered with lower-case letters, or, when working with a family of closely

related homologs, variable regions can be highlighted as such by replacing letters identical to the reference with periods.

Summary plots: conservation, consensus and quantitative annotation. Annotation is important for navigation, in both flat diagrams and interactive systems, because it guides the eye toward 'important' regions of an alignment. MSA workbenches and most of the editing and analysis tools reviewed here allow the user to interactively annotate alignments (see below). However, practically all MSA visualizations include some form of automatically generated annotation, such as consensus lines and alignment quality plots, displayed either above or below the alignment. Consensus annotation has its roots in the textual alignment files generated by MSA programs, but a variety of plots are now provided by modern tools (**Fig. 3**). Quality and consensus plots are calculated from each column's symbol frequency distribution using one of the many measures available[14,15]. Alternatively, sequence logos[16,17] provide a user-friendly indication of the dominant symbols at each position of the alignment. As in shading, described above, annotation can result from other kinds of calculations. For example, PFAAT[18], MacVector[19], VectorNTI[3] and Geneious are able to compute and plot averaged physicochemical quantities such as isoelectric point, and STRAP[20] supports extension of the program to allow complete customization.

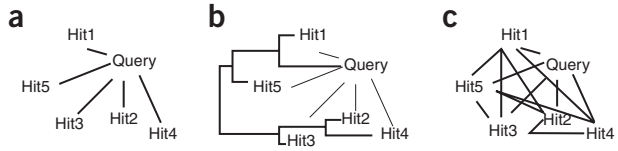

Figure 1 | Alignment topologies. Consistency graphs demonstrating complexity of different types of alignment algorithm. Nodes represent the query and each hit from the results of a sequence search, and edges indicate the mapping between a pair of sequences that the alignment algorithm optimizes. (**a**) MSA constructed directly from pairwise database search results. (**b**) MSA constructed using a guide tree, in which closely related sequences and then groups of sequences are optimally aligned. (**c**) MSA from consistency-based algorithms such as T-Coffee, in which all sequences are optimally aligned with one another.

Alignment editing, analysis and annotation

Integrated systems to support the editing and analysis of sequences have become possible with increased computing power and the ubiquity of internet connectivity. Most of the tools for MSA visualization mentioned here provide alignment coloring, shading and automated annotation facilities, as described above. However, 'editing and analysis' tools and most of the MSA workbench tools also allow alignments to be interactively edited,

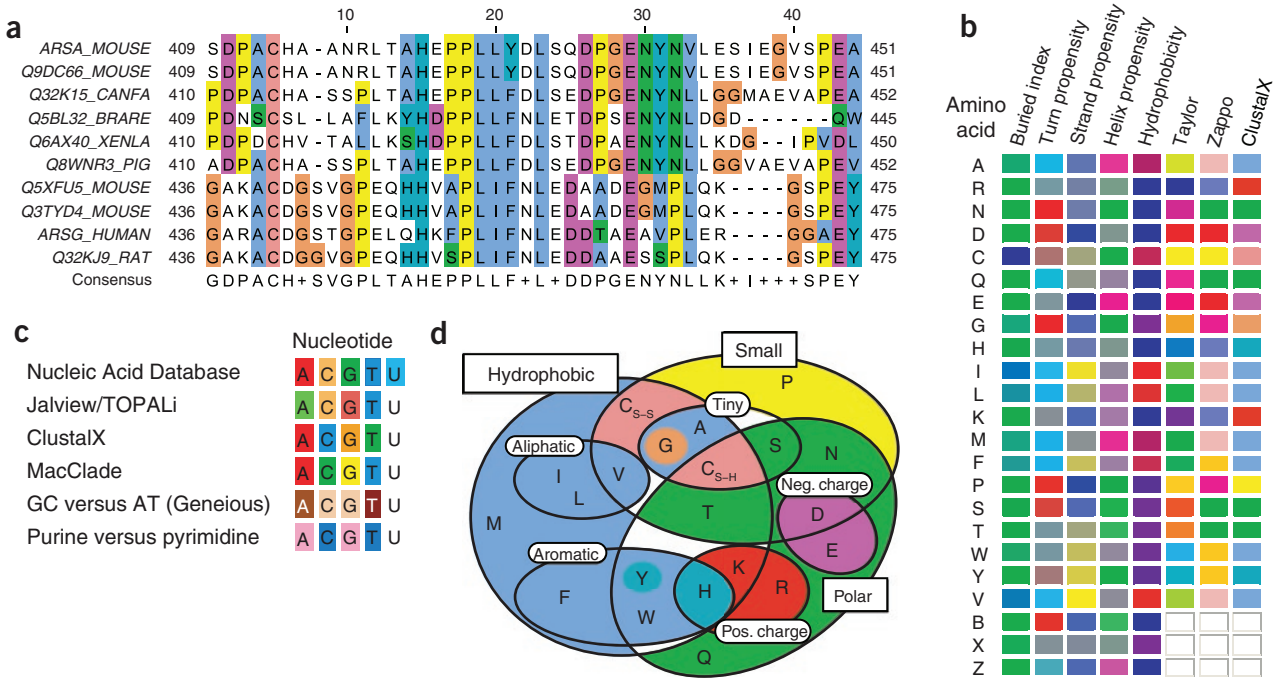

Figure 2 | Multiple alignment visualization. (**a**) A protein sequence alignment diagram rendered with Jalview[25]. Aligned sequences are arranged in rows and placed into a single reference frame, where each aligned position occupies a column in a table. Dashes indicate gaps. The label on the left-hand side of each sequence gives its Uniprot[53] entry name, start and end positions are shown at each end of its row and tick marks at the top allow a particular aligned column or sequence position to be read off. The consensus row at the bottom shows the most frequent residue at each column or a '+' if two or more residues are equally abundant. Residues in the alignment are colored according to the ClustalX[4] shading model: a color is only applied when that residue's abundance in the column is above a residue-specific threshold, highlighting potentially important residues (for example, proline and glycine) or patterns of conservation. (**b**) Examples of amino acid color schemes. Schemes are either quantitative, reflecting empirical or statistical properties of amino acids; or qualitative, reflecting an assignment according to physicochemical attributes. Zappo is a qualitative scheme developed by M. Clamp (personal communication); B, X and Z are amino acid ambiguity codes: B is aspartate or asparagine; Z is glutamate or glutamine; X is an unknown (or 'other'). (**c**) Examples of nucleotide color schemes used by the Nucleic Acid Database[54] and a selection of visualization tools. (**d**) Venn diagram after Taylor[55] showing the amino acids grouped according to their physicochemical properties. Coloring of each group (or amino acid label) is according to ClustalX, demonstrating the correspondence between color and physicochemical properties. Pos., positive; neg., negative.

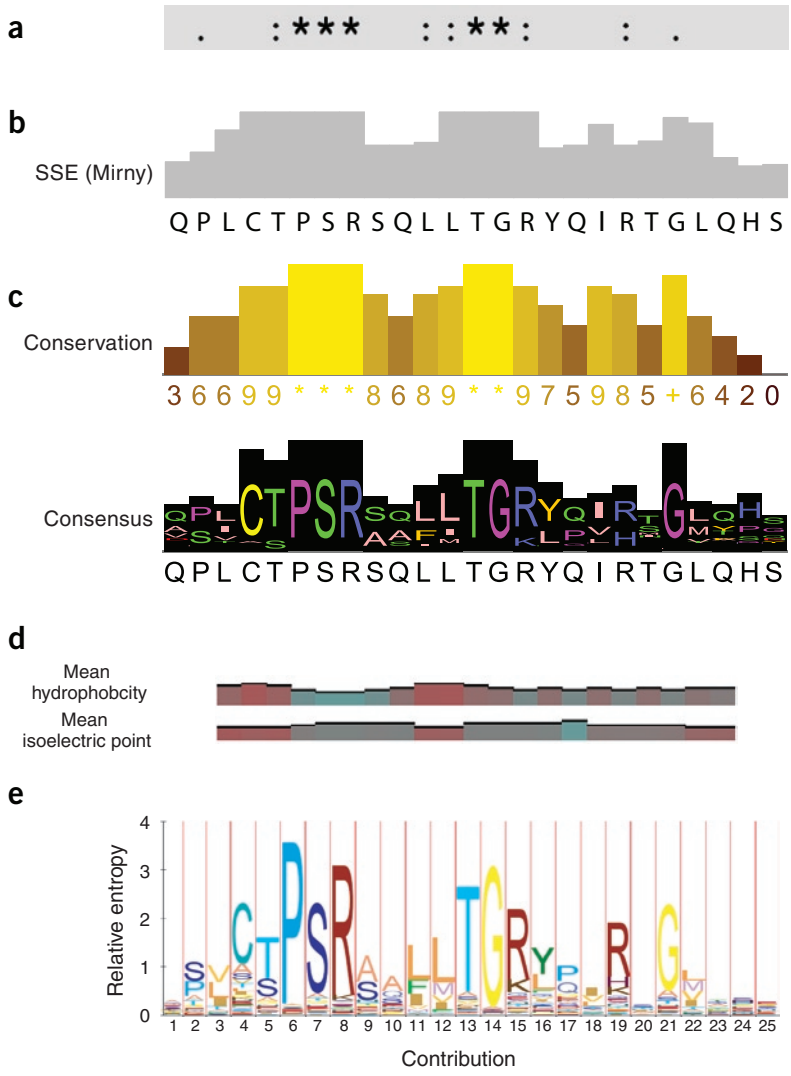

Figure 3 | Examples of automatically generated summary annotation for an alignment generated by MSA visualization tools. (**a**) ClustalW quality annotation from ClustalX: '*', ':' and '.' highlight identical, conserved and 'mostly conserved' columns, respectively, under a particular substitution model. (**b**) Mirny[56] conservation measure from PFAAT. Shannon entropy score is calculated for each column based on a reduced amino acid alphabet. (**c**) Amino acid physicochemical property conservation, consensus and overlaid sequence logo from Jalview. (**d**) Mean hydrophobicity and isoelectric point from Geneious. (**e**) HMMlogo visualization from Logomat-P (ref. 57) using corresponding HMMER[58] model. Labels have been added to the original images obtained from the tools in the creation of **b**,**d**,**e**.

annotated and analyzed, and often include additional visualizations for data associated with sequences or the result of analyses. Special support is provided for exploring sequence annotation, visualizing a sequence's associated protein or nucleic acid structure, or inspecting trees resulting from the application of phylogenetic analyses to the alignment. However, the degree of integrated visualization that these tools provide varies considerably. The latest tools use modern information visualization techniques, such as linked highlighting and brushing. For example, applying a color to a branch of a tree calculated from an MSA also colors the sequences in the linked MSA visualization. Conversely, older tools tend to provide either static or independent views of each type of data, but they often have unique visualization or analysis features; for example, GeneDoc has dedicated support for the MEME[21] motif discovery suite.

Editing and curation. MSAs from even the most accurate multiple alignment algorithm can contain errors, known as alignment artifacts[22], that make those regions of the alignment biologically meaningless. These occur because MSA algorithms find the optimal solution to a mathematical problem, rather than one reflecting the biochemical equivalence of the sequences. Such errors are hard to detect through automated means because correcting them often requires specialized knowledge of the protein or gene family. Editing and analysis tools such as Jalview are designed for alignment curation and so allow the user to modify parts of the alignment easily, either manually or with automated assistance[23], and allow changes to be undone. The shading and quality histograms in these tools also reflect changes immediately, to provide feedback on the effect of the modification.

Navigation, overviews, searching and selective row and column display. Systems such as PFAAT[18] and CINEMA[24] that are designed for curation, annotation and analysis provide navigation aids, including bird's-eye or overview windows that locate the visible region in its wider context. Search functions are also essential, and tools vary in capability, but typically they allow the user to locate and select sequence name, position or sequence pattern matches. Some tools (for example, Jalview[25]) also allow the user to create multiple views on the same alignment and to hide rows or columns, thus juxtaposing regions far apart, to aid in exploration and figure composition.

Interactive alignment annotation. Curation and figure generation require a flexible user interface for interactive alignment annotation. For example, a set of sequences in an alignment might be interactively grouped on the basis of the user's own knowledge, or regions corresponding to aligned domains or sequence motifs might be annotated so they are highlighted with a different rendering style. Alternatively, annotation tracks for the alignment may be added above or below the sequences containing colored labels and symbols to indicate potentially conserved properties such as protein secondary-structure elements or RNA secondary-structure contacts. Tools that support alignment column annotation display them in a similar fashion to the automated alignment summary annotation and usually allow the user to create and modify them interactively. Modern systems such as Jalview, PFAAT and CINEMA also provide a means to import and export annotation, and they offer 'project files' that store the complete state of an annotated alignment, enabling the user to return to it at a later date.

BOX 1 SOURCES OF ANNOTATION

Annotation, at the level of a complete sequence or for a given subsequence, can be obtained by importing a flat file (such as a GFF file; http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml) containing information associated with sequences in the alignment, or by remotely accessing bioinformatics databases (for example, UniProt, PDB, InterPro), either directly or using the Sequence Retrieval System (SRS). Recent tools also retrieve annotation by means of programmatic web services, such as the Distributed Annotation System (DAS)[76].

The structural and functional annotation of genomic and protein sequences has been the object of a large community effort in recent years. Although experimental evidence is clearly the most reliable source of annotation, it is unfeasible for the huge number of sequences available today. Therefore, automated or semi-automated systems, such as HAMAP[77] or MACSIMS[78], are being developed to transfer (in a controlled way) the known annotation from the characterized sequences in the databases to the uncharacterized ones. These systems assume that closely related sequences generally share a similar three-dimensional fold and often have similar functions. Under these assumptions, MSAs are analyzed to identify and annotate regions of the sequences sufficiently homologous to conserve structure and, perhaps, function.

A crucial aspect of publicly disseminated annotation is ensuring that structural and functional information associated with a sequence is 'machine readable'. As a consequence, ontologies and other structured vocabularies are being developed to represent the knowledge in a formal way. Widely used ontologies include the Gene Ontology[79,80], which describes gene products in terms of their associated biological processes, cellular components and molecular functions, the Sequence Ontology[81,82], describing the features and attributes of biological sequences and the NCBI organismal classification[83].

Visualization of annotation. Sequence annotation is an increasingly important part of alignment visualization, as it enables the user to rapidly identify key regions that should be curated, or inspected for variation. Annotation is available from a variety of sources (such as the Distributed Annotation System (DAS); see **Box 1**), and tools typically provide a means of importing annotation from flat files (for example, GFF or GenBank files), or automatically retrieving it by means of web services provided by databases or DAS annotation servers. Annotation associated with the complete sequence, such as its originating organism or biological function, can be shown adjacent to the sequence name (PFAAT, CINEMA) or in 'mouse-overs' (Jalview). Local annotation, such as domains, catalytic sites and protein secondary-structure elements, are rendered at their aligned positions variously as colored boxes, glyphs or other annotation, and any extra information provided by means of mouse-overs and embedded hyperlinks to web pages. Many sources of annotation indicate the provenance of their individual annotation records, and it is important to distinguish experimentally observed and predicted annotation in visualizations (**Supplementary Fig. 2**). Workbenches, and some of the editing and analysis tools (including PFAAT and Jalview), also allow the interactive creation and manipulation of sequence annotation, making these tools useful for sequence annotation curation.

Investigation of function. The analysis of sequences from diverse organisms is one of the most powerful ways to probe the structure and function of biological systems[26]. Most of the strategies for this include phylogenetic tree–based alignment analysis, which is discussed in the penultimate section of this review. However, some alignment visualization tools also support alternative approaches for functional site analysis. PFAAT and CINEMA can highlight regions of alignments that match sequence motif databases (for example, PROSITE[27] and TRANSFAC[28,29]) or, in the case of GENEDOC, *ab initio* motif discovery predictions[30–32]. The principal component techniques used in tools such as SeqSpace[33] (implemented in Jalview) and, more recently, pHMM[34] enable them to present a more abstract view of an MSA, by representing sequences and aligned positions as points on a two- or three-dimensional interactive scatter plot. Here, interactive brushing allows the user to locate and select a cluster of sequences or correlated sites and to view their locations in the linked view of the alignment.

Combined alignment and three-dimensional structure visualization. A linked molecular structure viewer enables exploration and interpretation of specific mutations in an MSA (**Supplementary Figs. 2b** and **3a**). Most of the tools capable of tree-based alignment analysis (discussed below) also allow alignment shading to be transferred to an associated protein structure[25,35–37].

RNA alignment visualization

Unlike proteins, the tertiary structure of an RNA macromolecule is almost solely determined by the pattern of nucleotide base pairs formed as it folds. RNA alignment visualization tools (**Table 1**) provide specialized shading and annotation models for investigating and highlighting the conservation of this secondary structure, and some, such as 4SALE, provide linked visualizations of the network of base pairs. However, RNA alignments still present problems, and new ways of representing these MSAs are being sought[38].

Sequence analysis workbenches

Sequence alignment workbenches differ from the other tools described here in that they provide a wide range of data management, analysis and visualization capabilities, of which alignment, sequence and phylogenetic analysis are just components. Because of this, they tend to separate sequence and sequence annotation editing from alignment visualization—unlike tools with their roots in MSA visualization, which deal with sequence and alignment annotation within a single context.

Semantic sequence annotation visualization. Workbench and MSA editing and analysis tools vary greatly in the way in which they render positional sequence annotation on alignments. However, all include some standard mapping between the graphical representation used and each type of annotation (for example, domains are

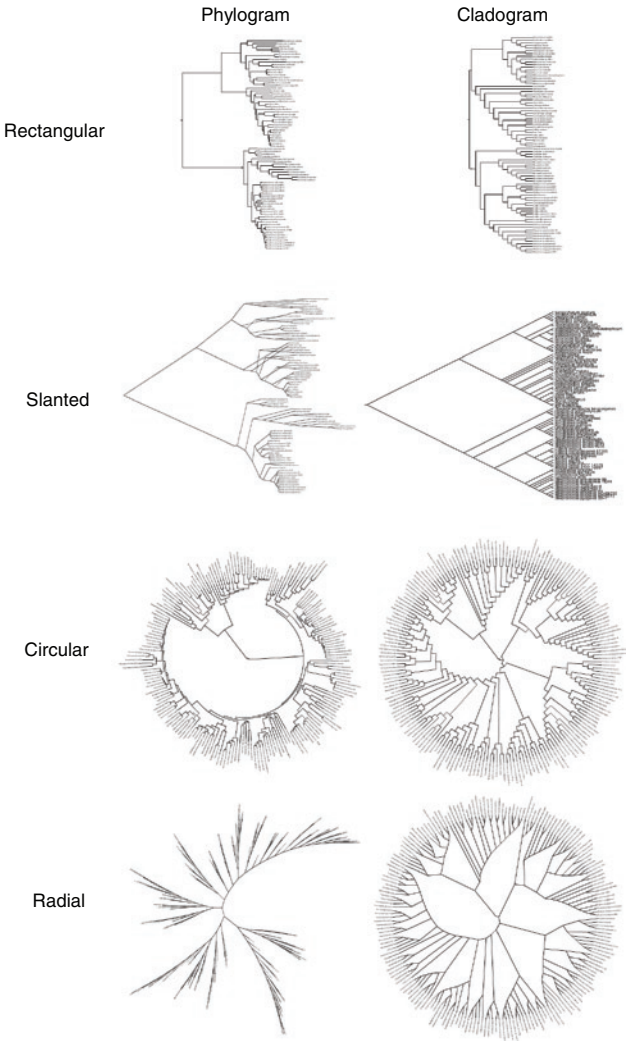

Figure 4 | Euclidean tree layouts. Trees are usually viewed as either phylograms, where branch length reflects similarity; or cladograms, where branch order reflects number of ancestors. Rectangular layouts highlight variation in branch length, whereas slanted layouts facilitate comparison of branch order (in cladograms). Circular layouts are most efficient when visualizing large numbers of taxa but make it more difficult to compare branch lengths. Radial layouts do not convey ancestral information and are most appropriate for unrooted trees, which are obtained when appropriate outgroups (reference phyla known to be less related to the phyla of interest than those are to one another) are not available.

displayed differently from metal binding sites). This enables them to exploit the formal terms found in annotation retrieved from public databases (**Box 1**) to display the annotation appropriately. Workbench tools with their roots in sequence visualization (such as MacVector and VectorNTI) provide the most advanced annotation display capabilities, and they allow each row of the MSA to be decorated with its own numbering, histograms, annotation tracks and complex diagrammatic glyphs.

VISUALIZING PHYLOGENETIC TREES

Phylogenetic analysis is an important part of the scientific workflow, and its backbone is the visual inspection, annotation and exploration of phylogenetic trees. It is therefore no surprise that the selection presented in **Table 2** barely represents the extensive repertoire of applications dedicated to tree visualization. These stand-alone or web-based tools can be placed into the three main functional classes used for MSA visualization tools, 'renderers', 'viewers', and 'workbenches', plus a further class, 'annotators'. Rather than detail the attributes of each class, we instead provide a basic guide to the principles underlying phylogenetic analysis (**Box 2**) and the tree visualization supported by these tools. We follow this by a brief discussion of the combined use of tree and MSA visualizations, and the state-of-the-art tools available for annotated phylogenetic visualization.

Basic tree terminology. Trees are directed graphs, in which branches connect internal (ancestral) nodes to their descendants. Leaf nodes represent elements of the character data used to construct the tree. These could be sequences that have been aligned in an MSA, or some other kind of characteristic information. In the case of MSA-based trees, two types are possible, which affects how the tree's internal structure should be formally interpreted. In gene trees, constructed from an MSA containing multiple gene families, internal nodes represent either speciation or gene duplication events. Conversely, if the MSA only contains sequences for a single gene in many species, the result is a species tree, and internal nodes then correspond to speciation events. However, precise interpretation of both types of tree requires knowledge of the wider evolutionary context, and specialized applications such as NOTUNG[39,40] have been developed to aid their analysis.

Tree visualization styles

Historically, phylogenetic trees were drawn to mimic real trees, from the ground up. However, with increasing numbers of nodes such tree layouts quickly become cluttered and difficult to read. Therefore, various alternative approaches are used to increase the readability and ease of annotation of trees, dependent on the tree size. These methods can be separated into two main categories, based on the geometry they use:

Euclidean geometry. This is the most common display method, and most tools in **Table 2** fall into this category. A variety of Euclidean tree styles exist (**Fig. 4**); but the choice of whether to present a cladogram or phylogram depends on the reliability of the evolutionary information available, and whether the tree is to be used to highlight differences in ancestry or rates of evolution. Trees with up to several hundred terminal nodes can be visualized with various rectangular layouts. Circular and radial layouts make it difficult to compare distal branches of the tree, but they are more useful for annotation, since they offer greater capacity for a given diagram size and can handle up to several thousand nodes.

Hyperbolic geometry. Hyperbolic display models are often used for very large network visualizations; tools that use this approach can easily handle thousands or even hundreds of thousands of nodes. Tools such as HyperTree[41], Hypergeny and Treebolic (**Table 2**) use hyperbolic projection to provide a view, analogous to that of a fish-eye lens, often termed 'focus+context'. This projection results in a circle in which distances between nodes of the tree are reduced exponentially, according to their distance from the center. By interactively panning the tree and bringing different branches to the central magnified region, it is possible to examine every part of the tree in detail while keeping a sense of the context. An

BOX 2 SOURCES OF PHYLOGENETIC DATA

Phylogenetic trees are calculated by applying mathematical models to infer evolutionary relationships between organisms, based on a set of characters that describe their differences. The most common characters are nucleotide or protein MSAs, but morphological information has also been used. There are four main categories of phylogenetic reconstruction methods: maximum parsimony, distance matrix, maximum likelihood and Bayesian approaches[84].

Parsimony is the principle of choosing simpler hypotheses in preference to those requiring a more complex explanation[85]. Maximum parsimony approaches create trees using the minimum number of ancestors needed to explain the observed characters[86].

Distance matrix methods, such as neighbor joining, allow more sophisticated evolutionary models than parsimony approaches. They estimate the mean evolutionary time (measured as the mean number of changes per site) since two species diverged from their most recent common ancestor[86]. However, because they reduce the estimate of most recent common ancestor to a single value, information on character evolution is lost.

Maximum likelihood and Bayesian methods constitute the state-of-the-art approaches for tree reconstruction. Maximum likelihood methods search a set of tree and evolutionary models to find the ones most likely to generate the observed characters[87]. Bayesian approaches offer more flexibility, as they allow optimization of all aspects of a tree (model, topology, branch length)[88]. But this comes at a cost: they require computationally expensive techniques such as Markov chain Monte Carlo to estimate terms in the Bayes equation.

alternative approach, introduced by H3Viewer, is to render trees in three dimensions embedded within a sphere, which allows the visualization of hundreds of thousands of nodes. A snapshot of a hyperbolic visualization of the whole NCBI taxonomy using Walrus[42] is shown in **Supplementary Figure 4**.

Tree-based alignment analysis

Phylogenetic trees and alignments are intrinsically related, and there are tools in **Tables 1** and **2** that can work with both kinds of data. However, tree-based alignment analysis[35,36,43–45] methods, which enable identification of functional motifs, are usually found only in MSA analysis tools. A notable example is Jevtrace[35] (**Table 1**; **Supplementary Fig. 5a**), which given an MSA and, optionally, an associated tree, partitions the aligned sequences into subfamilies, and automatically annotates columns containing sites showing variation significantly different from the overall tree. Other tools from **Table 1** that have interactive tree viewers usually allow the user to create manual or tree-based groupings on alignments. Once groups are defined, standard alignment shading models can highlight patterns of conservation and mutation that differ between groups (**Supplementary Fig. 5b**).

Annotation of trees

Phylogenetic trees in their raw form contain valuable information about the relatedness of the sequences (or other data) used to construct the tree, but it is possible to annotate the trees with further information, increasing their value for the interpretation of biological data. The most basic forms of annotation include branch lengths representing evolutionary distances and labels showing the phylogenetic support for each of the internal branches on the tree (such as bootstrap proportions). Tree branches are displayed in varying colors, either to highlight whole clades or to annotate particular features present in different nodes. Tools that support coloring of the tree branches include TreeDyn[46], Treevolution[47], iTOL[48] and FigTree. Some tools, such as TreeGraph[49], can also use the width of the nodes to convey quantitative annotation.

However, there is a growing need for tools capable of mapping more complex information onto trees. For example, metagenomic studies generate experimental results that are more easily interpreted when displayed using existing phylogenetic information. **Supplementary Figure 5** demonstrates such a visualization using the interactive Tree Of Life (iTOL), which allows users to annotate trees with various data set types, from simple histograms and pie charts to animated time series data and schematic representations of protein domain architectures. Such tools are already powerful, but their capabilities will need to be expanded further as phylogenetic trees become more commonly used in multidisciplinary investigation.

PERSPECTIVES AND CHALLENGES

Many alignment and tree visualization tools can display molecular structure and sequence annotation data, enabling in-depth analysis of a sequence family. Thanks to open standards and improved software and web services technology, more of these tools are becoming interoperable, and we expect them to provide increasingly flexible visualization and annotation interfaces for these types of biological data. Work has begun in this direction with the development of integrated tools such as eBIOX[11], Utopia[50] and general workbenches such as Bioclipse. In the future, we can expect integration with visualization tools for other kinds of 'omics' data, such as complete genome browsers and protein-protein interaction maps.

The ability to perform analyses and create annotation is an intrinsic property of tools designed for creating, editing and exploring trees and alignments. Systems such as DAS are now being exploited to facilitate the gathering of information from a host of bioinformatics databases, and it is possible to obtain rich and complex annotation derived from large-scale systems biology experiments. As a consequence, constraining the complexity of annotated visualizations is becoming necessary, calling for innovative visual representation techniques that aggregate and summarize annotations to make the most pertinent information accessible. Furthermore, biologists are notorious for the complex questions they ask of their data, and tools need more sophisticated query mechanisms to enable visualization of data selected on evolutionary and functional annotation criteria.

Lastly, other issues of scale must also be addressed. The sequence databases are growing exponentially, and the alignment of large sets of sequences has become a standard requirement. As an example, the largest protein family in the Pfam database contains over

100,000 sequences, but, given the accelerating rate of sequencing, it is likely that most families will contain thousands rather than hundreds of members within the next 5 years. Tools must therefore be improved to remedy any technical and conceptual limitations exposed when operating with such large data sets. Phylogenetic tools have already been developed that cope with relatively large trees (up to several thousand leaves), but size is still a particular problem for multiple alignment systems. Some show usability problems, such as poor interactive response times when loading or saving files or during other simple operations, such as selection or coloring. More generally, tools will need to provide access to the mass of information in these very large data sets, with enhanced overview displays that can summarize and provide easy navigation to more detailed views. In the case of trees, summary techniques include pruning and collapsing of branches. For sequence alignments, alternative visualization approaches such as partial order graphs[51] and circular alignment diagrams[52] have been developed, but, as far as we are aware, no interactive tool that supports them exists as yet. In conclusion, the increasingly dense biological data landscape presents new challenges for alignment and phylogenetic visualization. In response, exciting new approaches for the visualization and annotation of trees and alignments are being developed, and we look forward to using them in the future.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

J.B.P. acknowledges the support of the ENFIN European Network of Excellence (contract LSHG-CT-2005-518254) awarded to G.J.B. Several tools were made available as prereleases to the authors for evaluation purposes, and we thank the individuals and companies who obliged our requests.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
2. Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9 (2008).
3. Lu, G. & Moriyama, E.N. Vector NTI, a balanced all-in-one sequence analysis suite. *Brief. Bioinform.* **5**, 378–388 (2004).
4. Thompson, J.D., Gibson, T.J. & Higgins, D.G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* **2**, 2.3.1–2.3.22 (2002).
5. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
6. Edgar, R.C. & Batzoglou, S. Multiple sequence alignment. *Curr. Opin. Struct. Biol.* **16**, 368–373 (2006).
A comprehensive review of the approaches available for the alignment of many sequences.
7. Raghava, G.P., Searle, S.M., Audley, P.C., Barber, J.D. & Barton, G.J. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* **4**, 47 (2003).
8. Gouet, P., Robert, X. & Courcelle, E. ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* **31**, 3320–3323 (2003).
9. Barton, G.J. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* **6**, 37–40 (1993).
10. Goodstadt, L. & Ponting, C.P. CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics* **17**, 845–846 (2001).
11. Barrio, A.M., Lagercrantz, E., Sperber, G.O., Blomberg, J. & Bongcam-Rudloff, E. Annotation and visualization of endogenous retroviral sequences using the Distributed Annotation System (DAS) and eBioX. *BMC Bioinformatics* **10** (suppl. 6), S18 (2009).
12. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
13. Lin, K., May, A.C. & Taylor, W.R. Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types. *J. Theor. Biol.* **216**, 361–365 (2002).
The empirical analysis underlying the 'Taylor' amino acid color scheme; this builds on Taylor's earlier work (1986) concerning approaches for the classification of amino acids.
14. Valdar, W.S. Scoring residue conservation. *Proteins* **48**, 227–241 (2002).
15. Chakrabarti, S. & Lanczycki, C.J. Analysis and prediction of functionally important sites in proteins. *Protein Sci.* **16**, 4–13 (2007).
16. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
17. Schneider, T.D. Twenty years of Delila and molecular information theory: the Altenberg-Austin Workshop in Theoretical Biology biological information, beyond metaphor: causality, explanation, and unification Altenberg, Austria, 11–14 July 2002. *Biol. Theory* **1**, 250–260 (2006).
18. Caffrey, D.R. *et al.* PFAAT version 2.0: a tool for editing, annotating, and analyzing multiple sequence alignments. *BMC Bioinformatics* **8**, 381 (2007).
19. Rastogi, P.A. MacVector. Integrated sequence analysis for the Macintosh. *Methods Mol. Biol.* **132**, 47–69 (2000).
20. Gille, C. & Robinson, P.N. HotSwap for bioinformatics: a STRAP tutorial. *BMC Bioinformatics* **7**, 64 (2006).
21. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208 (2009).
22. Landan, G. & Graur, D. Characterization of pairwise and multiple sequence alignment errors. *Gene* **441**, 141–147 (2009).
To our knowledge, this is the first detailed analysis of the errors that may be introduced by tree based sequence alignment algorithms.
23. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12**, 543–548 (1996).
24. Lord, P.W., Selley, J.N. & Attwood, T.K. CINEMA-MX: a modular multiple alignment editor. *Bioinformatics* **18**, 1402–1403 (2002).
25. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
26. Margulies, E.H. & Birney, E. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat. Rev. Genet.* **9**, 303–313 (2008).
27. Hulo, N. *et al.* The 20 years of PROSITE. *Nucleic Acids Res.* **36**, D245–D249 (2008).
28. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* **9**, 326–332 (2008).
29. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
30. Zvelebil, M.J., Barton, G.J., Taylor, W.R. & Sternberg, M.J. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961 (1987).
31. Chakrabarti, S. & Panchenko, A.R. Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics* **10**, 207 (2009).
32. Horner, D.S., Pirovano, W. & Pesole, G. Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief. Bioinform.* **9**, 46–56 (2008).
33. Casari, G., Sander, C. & Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178 (1995).
34. Schwarz, R. *et al.* Detecting species-site dependencies in large multiple sequence alignments. *Nucleic Acids Res.* **37**, 5959–5968 (2009).
35. Joachimiak, M.P. & Cohen, F.E. JEvTrace: refinement and variations of the evolutionary trace in JAVA. *Genome Biol.* **3**, RESEARCH0077 (2002).
36. Goldenberg, O., Erez, E., Nimrod, G. & Ben-Tal, N. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.* **37**, D323–D327 (2009).
37. Li, W. & Godzik, A. VISSA: a program to visualize structural features from structure sequence alignment. *Bioinformatics* **22**, 887–888 (2006).
38. Brown, J.W. *et al.* The RNA structure alignment ontology. *RNA* **15**, 1623–1631 (2009).
39. Chen, K., Durand, D. & Farach-Colton, M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7**, 429–447 (2000).
40. Vernot, B., Stolzer, M., Goldman, A. & Durand, D. Reconciliation with non-binary species trees. *J. Comput. Biol.* **15**, 981–1006 (2008).

41. Bingham, J. & Sudarsanam, S. Visualizing large hierarchical clusters in hyperbolic space. *Bioinformatics* **16**, 660–661 (2000).
42. Hughes, T., Hyun, Y. & Liberles, D.A. Visualising very large phylogenetic trees in three dimensional hyperbolic space. *BMC Bioinformatics* **5**, 48 (2004).
43. Livingstone, C.D. & Barton, G.J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756 (1993).
44. Sankararaman, S. & Sjolander, K. INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics* **24**, 2445–2452 (2008).
45. Engelen, S., Trojan, L.A., Sacquin-Mora, S., Lavery, R. & Carbone, A. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput. Biol.* **5**, e1000267 (2009).
46. Chevenet, F., Brun, C., Banuls, A.L., Jacq, B. & Christen, R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, 439 (2006).
47. Santamaría, R. & Theron, R. Treevolution: visual analysis of phylogenetic trees. *Bioinformatics* **25**, 1970–1971 (2009).
48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
49. Müller, J. & Müller, K. TreeGraph: automated drawing of complex tree figures using an extensible tree description format. *Mol. Ecol. Notes* **4**, 786–788 (2004).
50. Pettifer, S. *et al.* Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinformatics* **10** (supp. 6), S19 (2009).
51. Raphael, B., Zhi, D., Tang, H. & Pevzner, P. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Res.* **14**, 2336–2346 (2004).
Introduces the partially ordered alignment algorithm and demonstrates how this graph based alignment visualization provides a more compact view of complex alignments.
52. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
Describes the CIRCOS approach for visualization of comparative genomic data, which can provide a more compact view of large multiple sequence alignments.
53. UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **37**, D169–D174 (2009).
54. Berman, H.M. *et al.* The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **63**, 751–759 (1992).
55. Taylor, W.R. The classification of amino acid conservation. *J. Theor. Biol.* **119**, 205–218 (1986).
56. Mirny, L.A. & Shakhnovich, E.I. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196 (1999).
57. Schuster-Böckler, B. & Bateman, A. Visualizing profile-profile alignment: pairwise HMM logos. *Bioinformatics* **21**, 2912–2913 (2005).
58. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
59. Seibel, P.N., Muller, T., Dandekar, T. & Wolf, M. Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Res. Notes* **1**, 91 (2008).
60. Wilm, A., Linnenbrink, K. & Steger, G. ConStruct: improved construction of RNA consensus structures. *BMC Bioinformatics* **9**, 219 (2008).
61. Jossinet, F. & Westhof, E. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **21**, 3320–3321 (2005).
62. Andersen, E.S. *et al.* Semiautomated improvement of RNA alignments. *RNA* **13**, 1850–1859 (2007).
63. Gille, C. Structural interpretation of mutations and SNPs using STRAP-NT. *Protein Sci.* **15**, 208–210 (2006).
64. Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. & Overington, J.P. JOY: protein sequence-structure representation and analysis. *Bioinformatics* **14**, 617–623 (1998).
65. Zmasek, C.M. & Eddy, S.R. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* **17**, 383–384 (2001).
66. Archer, J. & Robertson, D.L. CTree: comparison of clusters between phylogenetic trees made easy. *Bioinformatics* **23**, 2952–2953 (2007).
67. Huson, D.H. *et al.* Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**, 460 (2007).
68. Perrière, G. & Gouy, M. WWW-query: an on-line retrieval system for biological sequence banks. *Biochimie* **78**, 364–369 (1996).
69. Hillis, D.M., Heath, T.A. & St. John, K. Analysis and visualization of tree space. *Syst. Biol.* **54**, 471–482 (2005).
A demonstration of different kinds of tree visualization, and an examination of how spatial techniques such as multidimensional scaling can be used to visualize and compare ensembles of trees.
70. Page, R.D. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357–358 (1996).
71. Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L. & Zhou, Y. TreeJuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. *ACM Trans. Graph.* **22**, 453–462 (2003).
72. Kumar, S., Nei, M., Dudley, J. & Tamura, K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief. Bioinform.* **9**, 299–306 (2008).
73. Huson, D.H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
Describes the phylogenetic network visualization approach implemented in SplitsTree4, where evolutionary distance and bootstrap support are represented in one network structure, rather than an annotated tree.
74. Milne, I. *et al.* TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics* **25**, 126–127 (2009).
75. Jordan, G.E. & Piel, W.H. PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics* **24**, 1641–1642 (2008).
76. Prlić, A. *et al.* Integrating sequence and structural biology with DAS. *BMC Bioinformatics* **8**, 333 (2007).
77. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
78. Thompson, J.D. *et al.* MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics* **7**, 318 (2006).
79. Barrell, D. *et al.* The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.* **37**, D396–D403 (2009).
80. The Gene Ontology's Reference Genome Project. A unified framework for functional annotation across species. *PLoS Comput. Biol.* **5**, e1000431 (2009).
81. Reeves, G.A. *et al.* The Protein Feature Ontology: a tool for the unification of protein feature annotations. *Bioinformatics* **24**, 2767–2772 (2008).
82. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
83. Sayers, E.W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15 (2009).
84. Holder, M. & Lewis, P.O. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**, 275–284 (2003).
85. Swofford, D.L., Olsen, G.J., Waddell, P.J. & Hillis, D.M. Phylogenetic inference. in *Molecular Systematics* (eds. Hillis, D.M., Moritz, C. & Mable, B.K.) 407–514 (Sinauer, Sunderland, Massachusetts, USA, 1996).
86. Felsenstein, J. *Inferring Phylogenies* (Sinauer, Sunderland, Massachusetts, USA, 2004).
87. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
88. Huelsenbeck, J.P., Ronquist, F., Nielsen, R. & Bollback, J.P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314 (2001).