DOI: 10.1007/s10535-016-0649-8 BIOLOGIA PLANTARUM 60 (4): 619-627, 2016

REVIEW

A plant biologists´ guide to phylogenetic analysis of biological macromolecule sequences

F. CVRČKOVÁ*

Department of Experimental Plant Biology, Faculty of Sciences, Charles University, Viničná 5, CZ 128 43 Prague, Czech Republic

Abstract

Phylogenetic analysis has become a common step in characterization of gene and protein sequences. However, despite the availability of numerous affordable and more-or-less intuitive software tools, construction of biologically relevant, informative phylogenetic trees remains a process involving several critical steps that are inherently non-algorithmic, *i.e.*, dependent on decisions made by the user. These steps involve, but are not limited to, setting the aims of the phylogenetic study, choosing sequences to be analyzed, and selecting methods employed in sequence alignment construction, as well as algorithms and parameters used to construct the actual phylogenetic tree. This review aims towards providing guidance for these decisions, as well as illustrating common pitfalls and problems occurring during phylogenetic analysis of plant gene sequences.

Additional key words: bioinformatics, evolution, phylogenetic tree, protein domain identification, sequence alignment, sequence database searching.

Introduction: the "how" *versus* the "why"

Phylogenetic analysis became a standard part of the biologists' methodological toolbox, up to the point of being practically compulsory in studies dealing with characterization of new genes, including those from plants. With the proliferation of affordable, often free software, such as the continuously evolving *PHYLIP* (phylogeny inference package) toolbox (Felsenstein 1989), *PAUP* (phylogenetic analysis using parsimony; Wilgenbusch and Swofford 2003) or the intuitive and easy to use *MEGA* (molecular evolutionary genetics analysis) package (Hall 2013, Tamura *et al.* 2013), and even user-friendly web based tools represented, *e.g.*, by the *Phylogeny.fr* (Dereeper *et al.* 2008), *Phylemon* (Sánchez *et al.* 2011), or *T-REX* (tree and reticulogram reconstruction; Boc *et al.* 2012) servers, phylogenetic tree construction became seemingly easy even for biologists with little or no background in evolutionary biology, bioinformatics, or advanced computing. A beginner phylogeneticist is free to explore a jungle of possible ways leading from posing an initial question through data collection and analysis to a (more or less) biologically meaningful interpretation of results. Like in a real-world jungle, it is rather easy to become lost.

Several excellent guides focused mainly on the practical and technical aspects of phylogenetic tree construction, including specific instructions for use of

Submitted 9 December 2015, *last revision* 10 March 2016, *accepted* 12 April 2016.

Abbreviations: *BLAST* - basic local alignment search tool; CD-search - conserved domain search; *COBALT* - constraint-based multiple protein alignment tool; *DDBJ* - DNA data bank of Japan; *ENA* - European nucleotide archive; *INSDC* - international nucleotide sequence database collaboration; *MACAW* - multiple alignment construction and analysis workbench; *MAFFT* - multiple alignment using fast Fourier transform; *MEGA* - molecular evolutionary genetics analysis; ML - maximum likelihood; *MUSCLE* - multiple sequence comparison by log-expectation; *NCBI* - National Centre for Biotechnology Information; NJ - neighbor-joining; *PAUP* - phylogenetic analysis using parsimony; *PHYLIP* - phylogeny inference package; *SMART* - simple modular architecture research tool; *T-REX* - tree and reticulogram reconstruction.

Acknowledgements: I thank the many generations of students of my Introduction to Bioinformatics undergraduate course for providing continuous feedback that helped to shape the ideas presented here, Anton Markoš, Vojtěch Žárský and Shigehiro Kuraku for critical reading of this manuscript, and the Ministry of Education of the Czech Republic for financial support from the NPUI LO1417 project.

* Address for correspondence; e-mail: fatima@natur.cuni.cz

commonly used free software and public databases, have been published (*e.g.*, Baldauf 2003, Hall 2013, O'Halloran 2014), at least one of them specifically aimed at plant biologists (Harrison and Langdale 2006). A very recent article by Kuraku *et al.* (2016) concentrates mainly on the use of molecular phylogenies in reconstructing temporal sequences of events in the context of metazoan evo-devo biology but presents many important general theoretical insights, and can thus be recommended also to readers outside the animal biology field. The present paper aims towards providing an up-to-date "walkthrough" of some of the many possible paths that can produce a meaningful phylogenetic tree from protein or nucleotide sequences, with a specific focus on biological questions and theoretical and strategic decisions faced in the course of such analysis, *i.e.*, the "why" rather than the "how" of the procedure, albeit some technicalities are inevitable. Common problems and pitfalls will be also addressed where appropriate and illustrated by examples, often based on author's previous work.

What kind of questions can phylogenetic analysis solve?

Phylogenetic analysis produces a "tree" – a diagram that orders and connects the entities under study (in our case biological sequences) according to their (present-day) shared characteristics in a manner that is commonly interpreted as reflecting their mutual genealogical relationships. The sequences are presented as end nodes of a branched diagram whose inner nodes correspond to hypothetical shared ancestors. Branch lengths reflect the degree of diversification, and binary branching is universally expected, albeit conventional graphical representation may obscure either of these assumptions (Fig. 1; see also Baum 2008, Kuraku *et al.* 2016). Typically, we presume that the sequences are mutually homologous (a qualitative characteristic, meaning that they are derived from a common ancestor), and that they therefore share a certain degree of similarity (a quantitative characteristic).

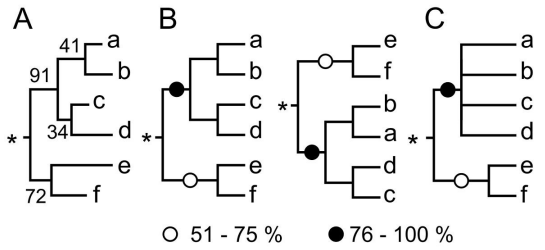

Fig. 1. Different formally correct graphical representations of the same phylogenetic tree constructed from six sequences (a to f) with position of the hypothetical root (outgroup) marked by an *asterisk*. *A* - A full tree including branch lengths and bootstrap support percentage as a measure of branch reliability. *B* - Two equivalent tree topologies with symbols at branches denoting bootstrap support; note that bootstrap values under 50 % are not shown and relative branch length information has been discarded. *C* - A tree similar as in *B* but with poorly supported branches collapsed. Note that the tree in *A* carries the most information available.

Thus, an evolutionary narrative in the most general sense (*i.e.*, stating that the analyzed sequences have evolved from a common ancestor through "descent with modifications") is not the outcome of phylogenetic analysis, but rather its central assumption. Holding this assumption, we interpret results of phylogenetic analysis as evidence for (or against) hypotheses on specific evolutionary narratives concerning our genes or proteins of interest, for instance answering the question whether a particular isoform of our protein arose before or after the separation of monocots and dicots. Phylogenetic analysis can "only" establish relationships, or similarities, within a set of entities on the basis of their shared and disparate characteristics. We interpret these relationships as being due to shared heritage, rather than to engineering skills or artistic antics of an external Creator, be he or she Almighty or human. Phylogenetic methods can be applied, in principle, to any set of objects which share common characteristics – even acquired at random or by convergence (Rieppel 2010). They have already been used also to analyze entitles where the "descent with modifications" narrative in the biological sense does not hold, such as literary texts or human-made artifacts, though even those can be understood as reflecting evolution in the realm of ideas, or memes (Howe and Windram 2011).

This is not a mere theoretical concern. Even if we stay firmly rooted in the evolutionary paradigm, we should be aware that phylogenetic analysis on its own cannot detect unintentional inclusion of data outside the studied set, that is sequences that do not share the assumed common evolutionary history with the rest of the group studied, due to, *e.g.*, sequence contamination in the source databases (Pible and Armengaud 2015), horizontal gene transfer, misidentification of relevant sequence portions, or unintended inclusion of distant paralogs (more on this in the sections on multidomain proteins and large gene families below). Data contamination does happen, and it is difficult to detect. An unexpected tree topology that would, for example, place an allegedly insect gene inside a group of exclusively plant sequences, or that would suggest, *e.g.*, the presence of a particular deep branch of the studied gene family only in a single species, may be an indicator of contamination. Suspected sequences should be carefully checked (see below) and excluded, or trimmed only to their relevant domains, if necessary, prior to re-calculation of the phylogenetic tree. With larger and more complex projects, the whole process of tree construction may thus develop into an iterative exercise, with repeated rounds of going back and forth between calculating and examining preliminary trees and optimizing the sequence set (and alignment).

Study design and input data acquisition

A typical aim of phylogenetic analysis is determining the relationship between a newly acquired sequence to its previously characterized homologs, *i.e.*, placing the new sequence into the context of existing knowledge on the gene family under study. Being well acquainted with the current state of knowledge is a self-obvious imperative, which should be also reflected in the choice of data for phylogenetic analysis. Inclusion of sequences that have been used previously to establish the basic topology of the gene family of interest is recommended. This should typically involve a full inventory in standard model angiosperm species with good quality genome sequence information available for at least *Arabidopsis thaliana* (L.) Heynh. and *Oryza sativa* L. var. *japonica*, possibly together with representative members of additional major angiosperm lineages and/or plants where the studied gene family has been characterized experimentally. Inclusion of non-angiosperm model plants, such as *Physcomitrella patens* (Hedw.) Bruch & Schimp. and *Selaginella moellendorffii* Hieron., or non-plant data serving as an outgroup, helps finding the root of the phylogenetic tree. This is especially important in studies aiming to reconstruct deep evolutionary history of the family of interest, a task that may grow very complex if involving large gene families, multidomain proteins, or even a combination of both (see below) and should, at best, be attempted after gaining some prior experience with simpler projects. However, inclusion of an outgroup (*i.e.*, a sequence that is related to all the members of the set studied, allowing its reliable alignment, but which itself is not part of the set) can be useful in all circumstances, as rooted trees are usually easier to interpret than unrooted ones. Nevertheless, an unrooted tree may be better than a rooted one constructed using a bad outgroup, leading to a substantial degradation of the sequence alignment. The outgroup can be, of course, added to a project, and tree topology re-calculated at any point.

When dealing with a single-copy gene whose evolutionary history never involved duplication, we would expect a tree merely reflecting the (usually already well-characterized) relationships between the organisms whose sequences were sampled. However, extant angiosperm genomes have undergone multiple rounds of whole-genome or chromosome segment duplication with subsequent paralog diversification and loss of some gene copies (Soltis *et al.* 2009). The presence of multiple paralogs of the gene studied in at least some of the organisms sampled is thus to be expected, and presents major challenges in the case of large gene families (see below). However, some genes do remain in single-copy due to selection pressure or chance, though often only in some species (Jiao and Paterson 2014; for some examples see, *e.g.*, Cvrčková *et al.* 2012). Even when analyzing a small gene family (with typical paralog numbers up to three), we should always aim towards obtaining a full inventory of relevant sequences in all the genomes sampled.

If dealing with protein-coding genes, phylogenetic analysis should be performed on the predicted protein product sequences, at least initially. Due to the larger amount of permitted characters in the sequence string (20 amino acids compared to 4 DNA bases), amino acid sequence-based searches are more sensitive than those utilizing only DNA information. This includes also searches for shared sequence motifs during construction of multiple alignments. Moreover, the existence of synonymous codons and variable 3rd codon positions means that over a third of DNA sequence may be randomized while maintaining the sequence of its protein product unchanged. With somewhat divergent protein products of, *e.g.*, 25 % protein sequence identity, still biologically relevant if occurring over several hundreds of amino acids (Chothia and Lesk 1986), the level of DNA sequence similarity may drop below detection limits. Thus, protein sequences are easier to find and align and are usually also more informative in subsequent analyses than nucleotide ones. The only exceptions are near-identical protein sequences which do not carry enough informative differences on the amino acid level. In such cases, phylogenetic analysis should be performed in parallel on both nucleotide and protein sequences to increase its sensitivity (Dvořáková *et al.* 2007), utilizing tools such as *BioEdit* (Hall 1999) to perform nucleotide sequence alignment guided by the predicted protein sequences. Of course, in the case of DNA sequences that do not encode a protein product, the analysis has to be performed on the nucleotide sequences themselves.

Public sequence databases, such as *UniProt* (Bateman *et al.* 2015) or databases of the international nucleotide sequence database collaboration (*INSDC*, Cochrane *et al.* 2011), comprising the European nucleotide archive (*ENA*), GenBank, and the DNA data bank of Japan (*DDBJ*), including their protein translation sections, are the default source of sequence data. If dealing with well-characterized plant genomes, the "reference sequences" section of these databases should be the first resource consulted. For information on plant genomes, specialized sites, such as the *U.S. Department of Energy Phytozome* database (Goodstein *et al.* 2012) or *Gramene* (Monaco *et al.* 2014), as well as resources dedicated to particular taxa such as *SolGenomics* (Fernandez-Pozo *et al.* 2015), are also worth visiting.

While keyword searches may yield sequences that can be used to initiate systematic searches for homologs, their results are bound to be affected by a subjective bias because they ultimately depend on the (very variable) quality of database sequence annotation. Sequence-based search tools, such as *BLAST* (basic local alignment search tool; McGinnis and Madden 2004, Johnson *et al.* 2008), are thus the main method for collecting input data, and the only reliable technique for exhaustively identifying homologous sequences. It is important to realize that the ability of *BLAST* to report divergent but still significantly related sequences depends inversely on the size of the

database searched. In the presence of many sequences very closely related to our query, some more distant but still relevant subjects may end up below the reporting threshold. Searching taxonomically restricted databases (*e.g.*, *Viridiplantae*, or even a single species of interest, in the *NCBI* implementation of *BLAST*) helps avoiding such a loss of relevant information.

Dealing with large gene families

Many plant genes are members of large gene families with dozens or even hundreds of paralogs, which makes phylogenetic analyses challenging (for a few examples see Dvořáková *et al.* 2007, Grunt *et al.* 2008, Yuksel and Memon 2008, Gish and Clark 2011, Cvrčková *et al.* 2012). To complicate things further, paralog numbers sometimes vary even among accessions or cultivars within species (Żmieńko *et al.* 2014). While construction of a complete phylogenetic tree involving all paralogs from all organisms studied would utilize maximum information in the available data, for very large families such a task is very demanding. If the overall structure of the gene family has already been described and we are "only" interested in putting a newly cloned gene or cDNA into evolutionary context, it is worth focusing the analysis only on a single branch of the gene family that contains our gene of interest, with representatives of more distant branches serving as outgroups. A detailed analysis of the whole gene family would be merely confirmatory in such a case, besides of being unnecessarily laborious.

Care has to be taken to avoid data contamination. Sensitive sequence-based searches by *BLAST* (and many other algorithms) will pick up sequences from distant branches of the protein family of interest, only loosely related to the query, and such sequences should not be included in the subsequent phylogenetic analysis except as outgroups, despite their good (*i.e.*, low) expected (E) values. For example, the enzyme phosphoglucomutase exists in plants in two isoforms, cytoplasmic and plastid-localized (Mühlbach and Schnarrenberger 1978), encoded by separate branches of the phosphoglucomutase gene family (Egli *et al.* 2010). A *BLAST* search for plant homologs of one of the *A. thaliana* cytoplasmic phosphoglucomutases (NP_177230) in the *Viridiplantae* section of the GenBank will yield around position 100 of the results list also the *A. thaliana* plastidic version (NP_199995) with the impressive E value of less than 10^{-150}. If we were to generate a phylogenetic tree of cytoplasmic phosphoglucomutases only, this sequence would be considered a contamination. A reverse *BLAST* search with the suspected sequence as a query can help revealing such extraneous sequences. If something else than the original query comes out on top of the results list, the suspect is very likely a contamination. In our example, the plastidic phosphoglucomutase retrieves a host of plastidic isoforms from various species on the top of the list of *BLAST* results, clearly identifying this protein as a plastidic phosphoglucomutase.

Gene and protein terminology deserves a particular attention in phylogenetic projects dealing with large sequence families. While database accession numbers should always be presented, they are not very informative on their own, and it is thus useful to introduce short, human-readable sequence identifiers. This should be done in a considerate way facilitating discussion of previous works by others. If a previously established gene terminology exists, it should be used preferentially, although one should not feel forced to use terminology contradicting his or her phylogenetic findings. If results of a phylogenetic analysis disprove an already existing classification, a new terminology may be proposed but sequence identifiers previously used in the literature should be cited as well, and, most importantly, possible causes of disagreement with previous studies have to be discussed (see Eliáš *et al.* 2002). An exhaustive description of the methods used is an important prerequisite for such a discussion.

Dealing with multidomain proteins

An extreme care to avoid data contamination has to be taken also when studying proteins that may contain segments that do not share evolutionary history with the rest of the molecule. This is typical for multidomain proteins whose evolution has involved domain acquisition, multiplication, or loss. In plants, protein domain architecture is typically only partly conserved between species or higher taxa (Zhang *et al.* 2012; for examples see also Cvrčková *et al.* 2004, Grunt *et al.* 2008), further complicating phylogenetic analyses.

Unless the protein sequences under study can be readily and unambiguously aligned along most of their length, their domain structures ought to be determined prior to constructing the multiple alignent that will be used for tree calculation in order to ensure that only homologous sequences are included in the analysis (Fig. 2). The presence of previously characterized domains in a protein sequence can be inferred by tools such as *CD-search* (conserved domain search; Marchler-Bauer and Bryant 2004, Marchler-Bauer *et al.* 2015) that is routinely run as a part of the *NCBI BLAST* searches, *ScanProsite* (De Castro *et al.* 2006), or *SMART* (simple modular architecture research tool ; Letunic *et al.* 2015).

However, our protein might also contain domains that are not characterized yet, and therefore not included in existing readily searchable databases. Such domains can

only be discovered in local homology-based database searches, such as *BLAST*, or during the construction of a multiple sequence alignment itself, where they will appear as islands of mutual sequence similarity found in all or a subset of the sequences analyzed. While conventional multiple sequence alignment tools, such as those discussed in the following sections, may help discovering previously uncharacterized motifs shared by all or some of the sequences examined, a single successfully aligned domain is bound to lock the sequences together, thereby preventing detection of other local islands of homology if they happen to occur in a varying order among the sequences. In this respect, the *MACAW* (multiple alignment construction and analysis workbench) program (Schuler *et al.* 1991) provides a unique ability to search for local blocks of sequence similarity in any user-selected subset of the examined sequences, or even of their parts, allowing thus detection of local islands of similarity incompatible with the preexisting alignment of other domains (Fig. 2). Unfortunately, this program was last updated in 1995, and no newer practically useable software with equivalent functionality exists, at least under *Microsoft Windows* (however, the *MACAW* paper is still being regularly cited). Due to its age, the *MACAW* installation archive is incompatible with post-*XP* versions of *Microsoft Windows*. However, a simple re-packaging of the installation archive into other archive formats, which can be performed on an old machine, allows its setup and use on newer versions of *Windows* or under *Linux* running vine. Like any other method aiming towards production of an internally consistent multiple alignment, *MACAW* cannot actually align sequences in a manner that would require their fragmentation. However, while an alignment joining simultaneously all the significant sequence similarity blocks cannot be displayed or exported in a single output, the user can gain a good insight into the domain layout of the proteins examined during *MACAW*-aided exploration.

Regardless of the techniques used to detect conserved domains in the protein sequences of interest, only matching (*i.e.*, mutually homologous) domains should be taken into the construction of the final alignment and phylogenetic tree calculation. It is, however, recommendable to leave a short (about 10 - 25 residues) overhang in front and behind the selected domains when preparing the input data for the sequence alignment step. The beginning and end of a conserved domain might not be recognized exactly in the initial domain architecture analysis, and it is always easier to trim an existing multiple alignment than to add missing residues during multiple alignment assembly.

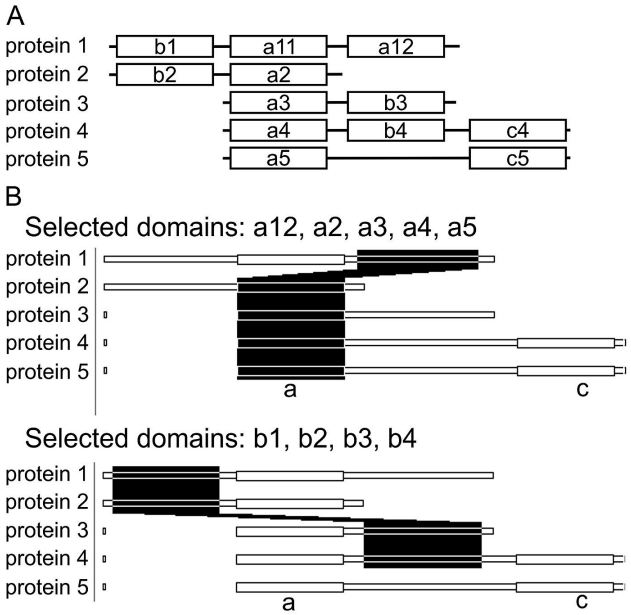

Fig. 2. A family of multidomain proteins. *A* - Localization of conserved domains (denoted by letters and numbers). Only domains with the same letters may be included together in a phylogenetic tree calculation. B - Screenshots of the *MACAW* program showing detection of repeated and swapped domains in protein sequences from *A*.

Constructing and processing multiple alignments

Numerous methods and programs for algorithmic multiple alignment construction are available, some of them as standalone applications, others as web-based services. For frequent alignment construction and computationally demanding large projects, locally installed programs should be used. Besides the classic *Clustal* (Higgins and Sharp 1988) and its various newer implementations (Larkin *et al.* 2007), other algorithms

such as, *e.g.*, *MAFFT* (multiple alignment using fast Fourier transform; Katoh 2013), *MUSCLE* (multiple sequence comparison by log-expectation; Edgar 2004), *T-Coffee* (Notredame *et al.* 2000), *Kalign* (Lassman *et al.* 2009), *COBALT* (constraint-based multiple protein alignment tool; Papadopoulos and Agarwala 2007) or *Dialign* (Al Ait *et al.* 2013), may be worth exploring, especially for divergent sequences where *Clustal* is known to perform poorly. However, all multiple alignment construction methods have some weaknesses that may lead to artifacts on certain problematic datasets, with definition of "problematic" varying somewhat unpredictably among algorithms. Generation of extraneous gaps is a common artifact of algorithmic multiple alignment construction.

Albeit systematic comparisons of various algorithms have been published (*e.g.*, Pais *et al.* 2014), and could be, in theory, used as a guidance for algorithm choice, such an approach is not very practical in real life. A reasonable alternative is simply performing the alignment by several methods and comparing their results. *M-Coffee*, a *T-Coffee* derivative, can construct a "meta-alignment" based on results of several algorithms (Moretti *et al.* 2007). Alignments can be also exported from almost any program in the commonly used *FASTA* format and then displayed side by side, *e.g.*, with the aid of *BioEdit* (Hall 1999), which is still being sporadically updated and whose latest (2013) version (available at http://www.mbio.ncsu.edu/bioedit/bioedit.html) is fully compatible with 64 bit versions of *MS Windows 7* and *8*. Their comparison can help identifying reliably aligned portions of the sequence (where the majority of algorithms agrees), but also provide alternative solutions for problematic areas (usually readily discernible by the presence of closely clustered gaps). For further work, a version of the alignment with fewest problematic sites should be chosen and subsequently manually modified where necessary (compare Blouin *et al.* 2009 for some good reasons), especially to resolve any gap clusters present.

Last but not least, prior to phylogenetic tree calculation, the alignment should be trimmed to remove any non-aligned or obviously non-homologous parts of the sequence, as well as regions containing (at least) large indels in more than one sequence. Optimally, all columns containing gaps in at least one sequence ought to be removed and only positions containing non-gap characters in all sequences should be taken into account in subsequent analysis, albeit this is not always possible. In some cases, sacrificing one or a few sequences that contain many or large gaps absent in the rest of the alignment may be a lesser evil than losing a substantial part of the alignment length. Phylogenetic tree calculation and validation algorithms involve, as a rule, the assumptions that *1*) amino acids at a given position are encoded by DNA sequences that have arisen from a common ancestor by point mutations and *2*) all positions within the alignment are equivalent. Removing indels as completely as possible ensures validity of assumption *1*. Removal of suspected misaligned sequences should also be as radical as feasible, following the rule "if in doubt, cut it out". While removing too much sequence might undermine the statistical significance of the tree obtained, it is unlikely to damage the result qualitatively if assumption *2* holds. However, retaining misaligned parts of the sequence would violate assumption *1*, leaving us with contaminated data.

While small alignments can be easily trimmed manually or with the aid of *BioEdit´*s "strip columns containing gaps" command, editing larger datasets can be made easier by using algorithmic tools such as *Gblocks* (Talavera and Castresana 2007), *trimAL* (Capella-Gutierrez *et al.* 2009), or *BMGE* (block mapping and gathering with entropy; Criscuolo and Gribaldo 2010).

Calculating, validating, and presenting phylogenetic trees

Several methods, based on very different theoretical approaches, can be used to cluster a set of previously aligned, mutually homologous sequences by their degree of similarity, *i.e.*, to produce a phylogenetic tree. One of the oldest, fastest, and still very frequently used is the neighbor-joining (NJ) method (Saitou and Nei 1987), quite satisfactory as a "quick and dirty" technique for gaining an initial insight into the tree topology, useful especially during the iterative process of optimizing the dataset composition and protein alignment. However, the NJ method is known to generate artifacts on datasets with highly variable distances (branch lengths), and therefore its results should be taken seriously and presented only if supported also by other methods. This means that a non-NJ method should always be employed in parallel or instead of NJ calculation, except, perhaps, in the rather exceptional cases where only closely related sequences are examined.

The maximum likelihood (ML) method (Goldman 1990) and Bayesian inference (Rannala and Yang 1996, Huelsenbeck *et al.* 2002) are the most commonly used methods for phylogenetic tree constructions free of the limitations of the NJ approach, although other techniques exist (for a good review with a detailed discussion of their strengths and weaknesses see Holder and Lewis 2003). Both methods share, however, the disadvantage of being computationally substantially more demanding than NJ. Rigorous application of the ML method including bootstrapping validation (see below) may make the required calculation times prohibitively long for large projects, albeit this problem can be overcome by using heuristic approximations such as, *e.g.*, *PhyML* (Guindon *et al.* 2010). While results of different tree construction methods applied to the same input data may look frighteningly disparate, it is important to realize that any tree branch can be freely rotated, and branch orientation

should be adjusted to highlight similarities when comparing trees (see Fig. 1*A*,*B*).

Regardless of the method used for tree construction, its results should never be presented without results of statistical validation that provide information on the reliability of individual branches. While Bayesian methods provide values of "posterior probability" together with the tree topology, the NJ and ML methods do not generate validation information on their own. Instead, validation of tree topology is commonly performed in a separate step by "bootstrapping", *i.e.*, by construction of typically several hundreds of trees from data sets derived from the input alignment by random sampling, and determination of the fraction of trees for which every particular branch of the original tree occurs. This fraction then determines the "bootstrap value" that can be interpreted as a measure of stability of the branch in question towards data perturbation. Bootstrapping can be applied also to Bayesian trees, and its results correlate rather well with posterior probabilities (Douady *et al.* 2003).

To improve readability of larger trees, low bootstrap values (below 50 %) can be omitted (with an appropriate comment in the figure legend); alternatively, symbols may be used instead of numeric bootstrap values (see Fig. 1*A,B*). Branches with low (below 50 %) support are also sometimes displayed collapsed, graphically resulting in a multifurcation (Fig. 1*C*). It is important to realize, however, that such a "multifurcation" only means an unresolved order of bifurcations. Moreover, such a display requires disposing the branch length information and its possibly biologically relevant interpretations, and it is questionable whether this is not too high a price for mere graphical highlighting an uncertainty in branch order.

Concluding remarks

The message of this paper may be summarized in a few rather oversimplified rules:

1) Good questions make good science. Good scientific questions are not only those which can be answered using the data and methods available, but those that also provide new insights into data, both new and old. Phylogenetic analysis projects should always be founded on a biologically meaningful question or hypothesis.

2) The well-established "rubbish in, rubbish out" rule holds. Like in any other area of science, the quality of the results of the final step depends on that of the input data and all preceding steps. If in doubt about a piece of data, throw it out.

3) Two methods are better than one. There is no single, objectively optimal method to conduct a phylogenetic study, construct a multiple alignment, or calculate a phylogenetic tree. Often there is no way to decide which method is better but trying several of them and comparing the results.

4) Construction of a multiple sequence alignment can be algorithmic but not objective. All algorithms rely on empirical parameters set by their authors in a way that should provide reasonable results with "typical" or "average" data. There is no warranty they will work for your data, and there is nothing wrong with exploring the space of possible parameter, or manually editing the alignment if necessary.

5) Know your data. Being familiar with a particular protein family means, for instance, being aware of functionally important sequence motifs. This is something algorithms cannot do since no software can critically evaluate published literature. Do not be afraid to use your theoretical background and experience when looking for domains or manually adjusting a protein sequence alignment. It also means being aware of existing knowledge, and projecting this awareness into terminology and interpretation of results.

References

Al Ait, L., Yamak, Z., Morgenstern, B.: DIALIGN at GOBICS – multiple sequence alignment using various sources of external information. - Nucl. Acids Res. **41**: W3-W7, 2013.

Baldauf, S.L.: Phylogeny for the faint of heart: a tutorial. - Trends Genet. **19**: 345-351, 2003.

Bateman, A., The uniprot consortium: UniProt: a hub for protein information. - Nucl. Acids Res. **43**: D204-D212, 2015.

Baum, D.: Reading a phylogenetic tree: the meaning of monophyletic groups. - Natur. Edu. **1**: 190, 2008.

Blouin, C., Perry, S., Lavell, A., Susko, E., Roger, A.J.: Reproducing the manual annotation of multiple sequence alignments using a SVM classifier. - Bioinformatics **25**: 3093-3098, 2009.

Boc, A., Diallo, A.B., Makarenkov, V.: T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. - Nucl. Acids Res. **40**: W573-W579, 2012.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T.: trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. - Bioinformatics **25**: 1972-1973, 2009.

Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. - EMBO J. **5**: 823-826, 1986.

Cochrane, G., Karsch-Mizrachi, I., Nakamura ,Y.: The international nucleotide sequence database collaboration. - Nucl. Acids Res. **39**: D15-D18, 2011.

Criscuolo, A., Gribaldo, S.: BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence

alignments. - BMC Evol. Biol. **10**: 210, 2010.
Cvrčková, F., Grunt, M., Bezvoda, R., Hála, M., Kulich, I., Rawat, A., Žárský, V.: Evolution of the land plant exocyst complexes. - Front. Plant Sci. **3**: 159, 2012.
Cvrčková, F., Pícková, D., Novotný, M., Žárský, V.: Formin homology 2 domains occur in multiple contexts in angiosperms. - BMC Genomics **5**: 44, 2004.
De Castro E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N.: ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. - Nucl. Acids Res. **34**: W362-365, 2006.
Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J,M,, Gascuel, O.: Phylogeny.fr: robust phylogenetic analysis for the non-specialist. - Nucl. Acids Res. **36**: W465-W469, 2008.
Douady, C.J., Delsuc, F., Boucher, .Y, Doolittle, W.F., Douzery, E.J.: Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. - Mol. Biol. Evol. **20**: 248-254, 2003.
Dvořáková, L., Cvrčková, F., Fischer, L.: Analysis of the hybrid proline-rich protein families from seven plant species suggests rapid diversification of their sequences and expression patterns. - BMC Genomics **8**: 412, 2007.
Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. - Nucl. Acids Res. **32**: 1792-1797, 2004.
Egli, B., Kölling, K., Köhler, C., Zeeman, S.C., Streb, S.: Loss of cytosolic phosphoglucomutase compromises gametophyte development in *Arabidopsis*. - Plant Physiol. **154**: 1659-1671, 2010.
Eliáš, M., Potocký, M., Cvrčková, F. Žárský, V.: Molecular diversity of phospholipase D in angiosperms. - BMC Genomics **3**: 2, 2002.
Felsenstein, J.: PHYLIP - phylogeny inference package (version 3.2). - Cladistics **5**: 164-166, 1989.
Fernandez-Pozo, N., Menda, N., Edwards, J.D., Saha, S., Tecle, I.Y., Strickler, S.R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., Mueller, L.A.: The sol genomics network (SGN) – from genotype to phenotype to breeding. - Nucl. Acids Res. **43**: D1036-D1041, 2015.
Gish, L.A., Clark. S.E.: The RLK/Pelle family of kinases. - Plant J. **66**: 117-127, 2011.
Goldman N.: Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. - System. Biol. **39**: 345-361, 1990.
Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S.: Phytozome: a comparative platform for green plant genomics. - Nucl. Acids Res. **40**: D1178-D186, 2012.
Grunt, M., Žárský, V., Cvrčková, F..: Roots of angiosperm formins: the evolutionary history of plant FH2 domain-containing proteins. - BMC Evol. Biol. **8**: 115, 2008.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. - System. Biol. **59**: 307-321, 2010.
Hall, B.G.: Building phylogenetic trees from molecular data with MEGA. - Mol. Biol. Evol. **30**: 1229-1235, 2013.
Hall, T.: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. - Nucl. Acids Symp. Ser. **41**: 95-98, 1999.
Harrison, C.J., Langdale, J.: A step by step guide to phylogeny reconstruction. - Plant J. **45**: 561-572, 2006.
Higgins, D.G, Sharp, P.M.: CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. - Gene **73**: 237-244, 1988.
Holder, M., Lewis, P.O.: Phylogeny estimation: traditional and Bayesian approaches. - Natur. Rev. Genet. **4**: 275-284, 2003.
Howe, C.J., Windram, H.F.: Phylomemetics – evolutionary analysis beyond the gene. - PLoS Biol. **9**: e1001069, 2011.
Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F.: Potential applications and pitfalls of Bayesian inference of phylogeny. - System. Biol. **51**: 673-688, 2002.
Jiao, Y., Paterson, A.H.: Polyploidy-associated genome modifications during land plant evolution. - Phil. Trans. Roy. Soc. London B Biol. Sci. **369**: 20130355, 2014.
Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L.: NCBI BLAST: a better web interface. - Nucl. Acids Res. **36**: W5-W9, 2008.
Katoh, K., Standley, C.M.: MAFFT multiple sequence alignment software version 7: improvements in performance and usability. - Mol. Biol. Evol. **30**: 772-780, 2013.
Kuraku, S., Feiner, N., Keeley, S.D., Hara, Y.: Incorporating tree-thinking and evolutionary time scale into develop-mental biology. - Dev. Growth Differentiation **58**: 131–142, 2016.
Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: Clustal W and Clustal X version 2.0. - Bioinformatics **23**: 2947-2948, 2007.
Lassmann, T., Frings, O., Sonnhammer, E.L.L.: Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. - Nucl. Acids Res. **37**: 858-865, 2009.
Letunic, I., Doerks, T., Bork, P.: SMART: recent updates, new developments and status in 2015. - Nucl. Acids Res. **43**: D257-D260, 2015.
Marchler-Bauer, A., Bryant, S.H: CD-Search: protein domain annotations on the fly. - Nucl. Acids Res. **32**: W327-W331, 2004.
Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Bryant, S.H.:CDD: NCBI's conserved domain database. - Nucl. Acids Res. **43**: D222-D226, 2015.
McGinnis, S., Madden, T.L.: BLAST: at the core of a powerful and diverse set of sequence analysis tools. - Nucl. Acids Res. **32**: W20-W25, 2004.
Monaco, M.K., Stein, J., Naithani, S., Wei, S., Dharmawardhana, P., Kumari, S., Amarasinghe, V., Youens-Clark, K., Thomason, J., Preece, J., Pasternak, S., Olson, A., Jiao, Y., Lu, Z., Bolser, D., Kerhornou, A., Staines, D., Walts, B., Wu, G., D'Eustachio, P., Haw, R., Croft, D., Kersey, P.J., Stein, L., Jaiswal, P., Ware, D.: Gramene 2013: comparative plant genomics resources. - Nucl. Acids Res. **42**: D1193-D1199, 2014.
Moretti, S., Armougom, F., Wallace, I.M., Higgins, D.G., Jongeneel, C.V., Notredame, C.: The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. - Nucl. Acids Res. **35**: W645-W648, 2007.
Mühlbach H, Schnarrenberger C.: Properties and intracellular distribution of two phosphoglucomutases from spinach

leaves. - Planta **141**: 65-70, 1978.

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.

O'Halloran, D.: A practical guide to phylogenetics for nonexperts. - J. visual Exp. **84**: e50975, 2014.

Pais, F.S.M., Ruy, P.C., Oliveira, G., Coimbra, R.S.:. Assessing the efficiency of multiple sequence alignment programs. - Algorithms mol. Biol. **9**: 4, 2014.

Papadopoulos, J.S., Agarwala, R.: COBALT: constraint-based alignment tool for multiple protein sequences. - Bioinformatics **23**: 1073-1079, 2007.

Pible, O., Armengaud, J.: Improving the quality of genome, protein sequence, and taxonomy databases: a prerequisite for microbiome meta-omics 2.0. - Proteomics **15**: 3418-3423, 2015.

Rannala, B., Yang, Z.: Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. - J. mol. Evol. **43**: 304-311, 1996.

Rieppel, O.: The series, the network, and the tree: changing metaphors of order in nature. - Biol. Phil. **25**: 475-496, 2010.

Sánchez, R., Serra, F., Tárraga, J., Medina, I., Carbonell, J., Pulido, L., de María, A., Capella-Gutíerrez, S., Huerta-Cepas, J., Gabaldón, T., Dopazo, J., Dopazo, H.: Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. - Nucl. Acids Res. **39**: W470-W474. 2011.

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

Schuler, G.D., Altschul, S.F., Lipman, D.J.: A workbench for multiple alignment construction and analysis. - Proteins **9**: 180-190, 1991

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., de Pamphilis, C.W., Wall, P.K., Soltis, P.S.: Polyploidy and angiosperm diversification. - Amer. J. Bot. **96**: 336-348, 2009.

Talavera, G., Castresana, J.: Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. - System. Biol. **56**: 564-577, 2007.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S.: MEGA6: molecular evolutionary genetics analysis version 6.0. - Mol. Biol. Evol. **30**: 2725-2729, 2013.

Wilgenbusch, J.C., Swofford, D.: Inferring evolutionary trees with PAUP*. - Current Protocols Bioinformatics **6**: Unit 6.4, 2003.

Yuksel, B., Memon, A.R.: Comparative phylogenetic analysis of small GTP-binding genes of model legume plants and assessment of their roles in root nodules. - J. exp. Bot. **59**: 3831-3844, 2008.

Zhang, X.C., Wang, Z., Zhang, X., Le, M.H., Sun, J., Xu, D., Cheng, J., Stacey, G.: Evolutionary dynamics of protein domain architecture in plants. - BMC Evol. Biol. **12**: 6, 2012.

Żmieńko, A., Samelak, A., Kozłowski, P., Figlerowicz, M.: Copy number polymorphism in plant genomes. - Theor. appl. Genet. **127**: 1-18, 2014.