

DRUIDS—Detection of Regions With Unexpected Internal Deviation From Stationarity

OLIVIER FEDRIGO,^{1*} DEAN C. ADAMS,² AND GAVIN J.P. NAYLOR³

¹*Department of Biology, Duke University, Durham, North Carolina 27708*

²*Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, Iowa 50011*

³*School of Computational Science and Information Technology 150-A Dirac Science Library, Florida State University, Tallahassee, Florida 32306*

ABSTRACT Most methods for inferring phylogenies from sequence data assume that patterns of substitution have been stationary over time. Changes in evolutionary constraint can result in nonstationary substitution patterns that are phylogenetically misleading unless modeled appropriately. Here we present a multiple-alignment-based method to identify regions that are likely to contain misleading phylogenetic signals due to changes in evolutionary constraints. The method uses a moving window approach to identify regions with a statistically significant deviation from stationarity in the physicochemical properties of amino acids among taxa. The protocol has been implemented in the software package DRUIDS (Detecting Regions of Unexpected Internal Deviation from Stationarity), available from the first author upon request. *J. Exp. Zool.* 304B:119–128, 2005.

© 2005 Wiley-Liss, Inc.

Phylogenetic trees are useful for understanding both evolutionary patterns and processes and are integral to many genomic and bioinformatic applications. They have been used to understand patterns of gene duplication (Page and Cotton, 2002), to estimate the function of unidentified genes using orthology (Storm and Sonnhammer, 2003), to identify residues responsible for functional divergence (Yokoyama and Radlwimmer, '98; Naylor and Gerstein, 2001; Gu and Vander Velden, 2002), and to identify subsets of amino acids under selection (Yang et al., 2000). As the number of disciplines embracing phylogenetic approaches continues to grow, it becomes ever more important to ensure that estimated trees are accurate. Empirical studies suggest there is considerable room for improvement as different markers frequently yield different trees when subjected to phylogenetic analysis.

The lack of agreement among phylogenetic estimates from different markers suggests that models of molecular evolution do not have a good fit to the data to which they are applied (horizontal transfer, lineage sorting, gene duplication, and sampling error notwithstanding). One of the more common shortcomings of models is their failure to accommodate deviations in process stationarity—changes in patterns of substitution associated with

changes in constraint over the course of evolution. Such changes in substitution patterns often show up as a lack of fit between the assumed model and the input data.

DETECTING REGIONS OF INCONSISTENCY BETWEEN MODEL AND DATA

Tree-based approaches

Grassly and Rambaut ('97) presented an approach (implemented in the software program PLATO: Partial Likelihood Analysis Through Optimization) to detect gene regions whose patterns of substitution had poor fits to a given tree for a specified model. Their motivation was to detect recombination events in gene sequences, but their method can be used to detect any region of relatively poor fit between data and model, whether it be due to a recombination event or to a change in constraint. PLATO receives a tree, a substitution model, and a DNA sequence align-

*Correspondence to: Olivier Fedrigo, Duke University, Department of Biology, 139 Biological Science Building, Science Drive, Durham, NC 27708-0338. E-mail: ofedrigo@duke.edu

Received 24 August 2004; Accepted 15 November 2004

Published online 31 January 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21032

ment and outputs the likelihood scores associated with specific regions of the molecule.

While it is effective for identifying regions that have a poor fit to a model for a particular tree, it requires a tree. As such, it can only be used to detect misleading signals when the tree is given.

To avoid the obvious circularity (requiring the tree to identify regions best suited to estimating the tree), a method is needed that identifies phylogenetically misleading regions without reference to a tree. If regions of nonstationarity could be so recognized, they could be deleted prior to analysis, or steps could be taken to develop appropriate models to better estimate trees.

Alignment-based (tree-independent) approaches

Among closely related taxa, it is common to find variation at the nucleotide level but not at the amino acid level. At higher levels of divergence, we begin to see variation in amino acids but not in the chemical properties of those same amino acids. At extreme levels of divergence, variability in chemical properties may exist among homologous regions of sequences as their functions begin to diverge. In general, when sequences show variation at the nucleotide or amino acid level but not at the level of chemical properties, we assume that functional constraints have been constant over the course of evolution. However, when homologous regions exhibit differences in chemical properties among taxa, we must entertain the possibility that there have been changes in functional constraints over the time scale being examined. Such changes in constraint can result in patterns of nucleotide and amino acid substitution that are nonstationary and potentially phylogenetically misleading unless modeled appropriately. We present an approach to identify such potentially misleading regions directly from the alignment. Because there is always some degree of “background variation” in natural systems, we identify regions whose variation in amino acid properties across taxa is higher than would be expected by chance. This allows us to distinguish variation due to deviation from stationarity from that due to poor taxon sampling and high rates of evolution that can also result in patterns of variation that are superficially similar to those resulting from a shift in constraint. We call the approach “Detecting Regions with Unexpected Internal Deviation from Stationarity,” or DRUIDS.

METHODS

Mapping higher-order features onto DNA sequences

In order to identify differences in higher order properties, we first assign mappings from the DNA to the corresponding amino acid sequence using the appropriate genetic code, and from the amino acid to the property associated with the amino acid (e.g., hydrophobic, acidic, basic, amide etc). Each amino acid is assigned a numerical score for hydrophathy, polarity, volume, mass, and charge using standard measures from the literature [e.g., the Kyte–Doolittle hydrophathy scale (K-D) (Kyte and Doolittle, '82) and residue volume (Zamyatin, '72)]. For example, if an adenine is encountered at a particular site for a particular taxon and this adenine occupies the second codon position of a “CAT” codon, coding for the hydrophilic residue histidine, we would assign the nucleotide the hydrophathy score of -3.2 (following the Kyte–Doolittle scoring system).

Assessing deviation from stationarity

To detect regional shifts in constraints, we assess the variation of each property across taxa using an overlapping sliding window approach (Fig. 1). Because the approach aims to detect

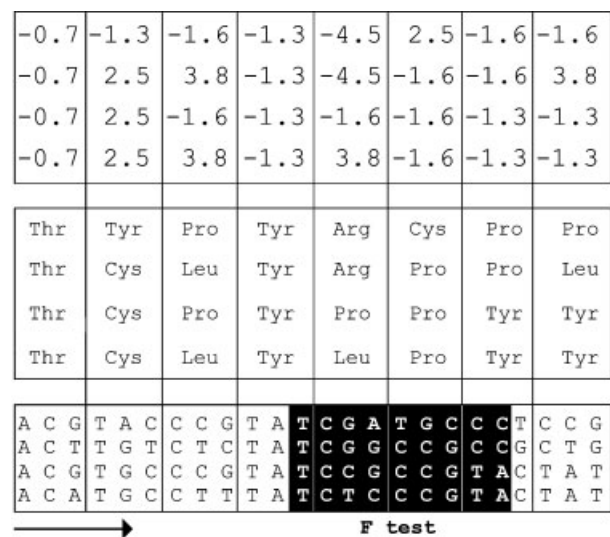


Fig. 1. Principle of the DRUIDS algorithm. A data set can be translated into amino acids and its corresponding values for a particular chosen property (e.g., the Kyte and Doolittle scale for hydrophathy). Values can be assigned to each nucleotide. Overlapping, sliding window, and F -test are calculated.

amino acid property patterns, the codon level seems to be the obvious level of analysis.

However, we chose a nucleotide-by-nucleotide overlapping sliding window approach to facilitate a direct comparison with other analysis such as G+C content and nucleotide substitution rate patterns. Comparative tests between the two levels showed a small difference at the edges of the detected region (data not shown) that is negligible relative to the effects of other approximations such as choice of a threshold of significance and window sizes. For each window, the null hypothesis is that there is no deviation from stationarity or “DFS” across taxa. This is achieved by contrasting the observed variation within a window across taxa, against a null distribution generated by resampling character states. The resampling procedure is based on 10,000 iterations and samples codons rather than nucleotides to conserve biological relevance and to be consistent with the property mapping. In all windows, the variance among taxa is used to calculate an F ratio.

The idea is to measure the variation among (MSb) and within (MSw) taxa. If the variation between taxa is significantly larger than the variation within taxa, we assume there is a deviation from stationarity inside the window.

$$F = MSb/MSw$$

with

$$MSb = \sum_{i=1}^k (\bar{\mu}_i - \bar{\mu})^2 / (k - 1),$$

which is the mean square between k taxa and with

$$MSw = \sum_{i=1}^k \sum_{j=1}^n (\mu_{ij} - \bar{\mu}_i)^2 / (N - k),$$

which is the mean square within taxa for a window of n nucleotides and $N=kn$; μ_{ij} denotes the value mapped on the j th nucleotide of the window for the i th taxon, $\bar{\mu}_i$ is the average of all nucleotides in the window for the i th taxon, and $\bar{\mu}$ is the mean of all nucleotides for all taxa (N).

Those F ratios that fall in the highest 5% or 10% of the distribution are taken as the subset most likely to exhibit a DFS. Using most extreme 5% of the F values ensures that our procedure has an experimental error rate of 5%.

Implementation

Software has been implemented for the Windows operating system and is in development

for Mac OSX. It is available by request from the first author or at <http://www.acpub.duke.edu/~ofedriego/Druids.html>. The DRUIDS software has a Graphical User Interface that allows one to explore up to three parameters simultaneously. A help file explaining the features and options is included. The user can select the properties to be investigated from the following: residue hydrophobicity, residue volume, residue mass, residue charge, and residue solubility. Additional properties can be added by placing a design file into the appropriate directory.

Local deviation from stationarity for G+C content can also be evaluated. Deviation from stationarity can be explored for multiple physical-chemical parameters (e.g., hydrophobicity and volume), but the alignment is scanned for DFS for each property separately. The union or intersection of the solutions for each parameter can then be explored depending on the goal of the study. The default value for window size is 12 nucleotides, corresponding to four amino acids, the approximate length associated with one turn of an alpha helix (3.6 residues long). The maximum is set at 21 nucleotides, which corresponds to seven amino acids or approximately two turns of an alpha helix. This setting may be useful to capture changes in constraint associated with a pair of residues located on one face of a helix (e.g., exposed to solvent). The smallest allowable window size is set at six nucleotides (two codons). When a range of window sizes is input, the program will return all the sites exhibiting a statistically significant deviation from stationarity in the final output. However, detection of patterns is done independently for each window size and/or property. There is no estimation of a global statistical significance. Other aspects of the alignment can also be explored, such as base composition, G+C content, and codon usage bias (Wright, '90; Long and Gillespie, '91).

Tests of the method

We evaluated the effectiveness of DRUIDS in three ways. We asked: (1) Can DRUIDS identify constraint change using simulated data? (2) Can DRUIDS identify regions of nonstationarity in real data sets? (3) Can DRUIDS be used to improve phylogenetic accuracy?

(1) Can DRUIDS identify constraint changes in simulated data?

We tested the efficacy of DRUIDS with simulated protein coding sequence data. Sequences were simulated over a known phylogeny (Fig. 2) using a codon model (Goldman and Yang, '94; Muse and Gaut, '94) in which each codon was allowed to mutate to one of its nine possible one-step neighbors (sequence size=1200 nucleotides and equal branch length=0.25 substitution per site). The freedom to vary of each codon was restricted according to the physical-chemical properties of the constraint class to which it belonged (i.e., codons constrained to be hydrophobic were allowed to mutate to other codons whose amino acids were also hydrophobic). The ancestral starting sequence included a patch at the 5' end (patch I) constrained to be hydrophobic (node 1, Fig. 2). After two bifurcation events (node 2, Fig. 2), the hydrophobic constraint in patch I was changed to a hydrophilic constraint (subsequently allowing only hydrophilic residues to be substituted). At the same time, a new hydrophobic constraint was introduced at a patch toward the 3' end of the sequence (patch II) for the clade ((D,E),(F,G)),((H,I),(J,K))). The swapping of the hydropho-

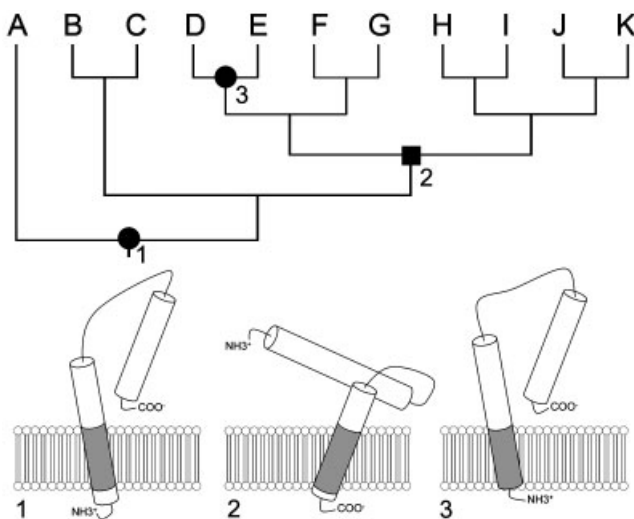


Fig. 2. Simulation design. Filled circles represent a hydrophobic constraint at the NH_3^+ end of the protein (corresponding to the 5' end of the gene). The square represents a hydrophobic patch at the COO^- end of the protein (corresponding to the 3' end of the gene). These two constraints alternate at nodes 1, 2, and 3 and are illustrated in their respective panels. This series of evolutionary changes in constraint were chosen to mimic the alternating movements of two helices from a transmembrane to an extramembrane environment.

bicity constraint for a hydrophilic constraint with the simultaneous establishment of a new hydrophobicity constraint was chosen to mimic the shifts in constraint associated with moving from a transmembrane to an extramembrane environment. Finally, the original hydrophobic constraint on patch I, at the 5' end of the sequence, is restored at the base of the clade containing taxa D and E (node 3, Fig. 2). The sequences that were simulated under this constraint were subjected to parsimony analysis and maximum likelihood with GTR model (Lanavé et al., '84) using PAUP* 4.0b10 (Swofford, '99), and the phylogenetic signal was characterized with split decomposition (Huson, '98). Sequence regions showing deviation from stationarity were identified and filtered out using DRUIDS with a window size of 12 and 5% significance level. The filtered data set was then reanalyzed using parsimony, maximum likelihood, and split decomposition.

(2) Can DRUIDS identify regions of nonstationarity in real data sets?

We subjected four different vertebrate cytochrome *b* data sets, each with similar numbers of taxa and comparable divergence times, to analyses with DRUIDS. The four data sets were (i) birds, (ii) carnivores, (iii) primates, and (iv) teleost fishes. Patterns of nonstationarity were compared among the data sets. These were then plotted onto the three-dimensional crystal structure of cytochrome *b* (Xia et al., '97) using the software package MolMol (Koradi et al., '96).

(3) Can DRUIDS be used to improve phylogenetic accuracy?

If regions of nonstationarity are responsible for some of the phylogenetically misleading signal in molecular data sets, then identifying and deleting them should improve phylogenetic estimation. We tested this idea using a data set composed of deeply divergent mitochondrial sequences from vertebrate groups whose evolutionary inter-relationships are well established (Fig. 5) but that are known to yield anomalous trees for mitochondrial sequence data (Rasmussen and Arnason, '99; Takezaki and Gobojoji, '99). This "known phylogeny" provided a benchmark against which phylogenetic estimates could be contrasted. We explored the effects of deleting nonstationary regions for three genes: cytochrome oxidase I (coI), cytochrome *b* (cytb), and NADH2 (nd2) (Table 1).

TABLE 1. Taxa list and GenBank accession number

GenBank	Latin name	Common name
NC.001567	<i>Bos taurus</i>	Cow
NC.001794	<i>Macropus robustus</i>	Wallaroo
AY235571	<i>Gallus gallus</i>	Chicken
NC.002785	<i>Struthio camelus</i>	Ostrich
Y13113	<i>Alligator mississippiensis</i>	Alligator
NC.001727	<i>Crossostoma lacustre</i>	Loach
NC.001717	<i>Oncorhynchus mykiss</i>	Trout
NC.001606	<i>Cyprinus carpio</i>	Carp
NC.000890	<i>Mustelus manazo</i>	Star spotted dogfish
NC.002012	<i>Squalus acanthias</i>	Spiny dogfish
NC.000893	<i>Raja radiata</i>	Starry skate
NC.001626	<i>Lampetra marinus</i>	Lamprey

Phylogenetic trees were estimated using parsimony and maximum likelihood with the GTR model for each gene individually and in combination. Each data set was then subjected to analysis using DRUIDS. Regions showing nonstationarity for both hydrophobicity and volume (union of the two properties for 21-nucleotide-long windows with 10% for the F -test significance) were removed by filtering. The “filtered” data sets were then subjected to parsimony and maximum likelihood analyses. The effects of these exclusion strategies were contrasted with the effects of removing third position sites, an alternative commonly used strategy to “improve” phylogenetic signal (Swoford et al., '96).

RESULTS AND DISCUSSION

Can DRUIDS identify constraint changes in simulated data?

Maximum parsimony retrieved the correct tree topology when applied to the simulated data for which there had been no change in constraint. It retrieved an incorrect topology (with bootstrap support values of 100%) when applied to the data set that had been subjected to changes in constraint (Fig. 3A). The split-tree analysis revealed conflicting signals for the (B,C,A,D,E) clade (Fig. 3C). The DRUIDS analysis correctly detected the areas for which the constraints differ, namely, positions 165–531 and 685–1140. These areas are, except for few nucleotides due to edge effect, almost identical to those imposed in the simulation (patch I, 168–533; patch II, 685–1141). The “true” phylogeny was obtained after removing these nonstationary regions (Fig. 3B). Similar

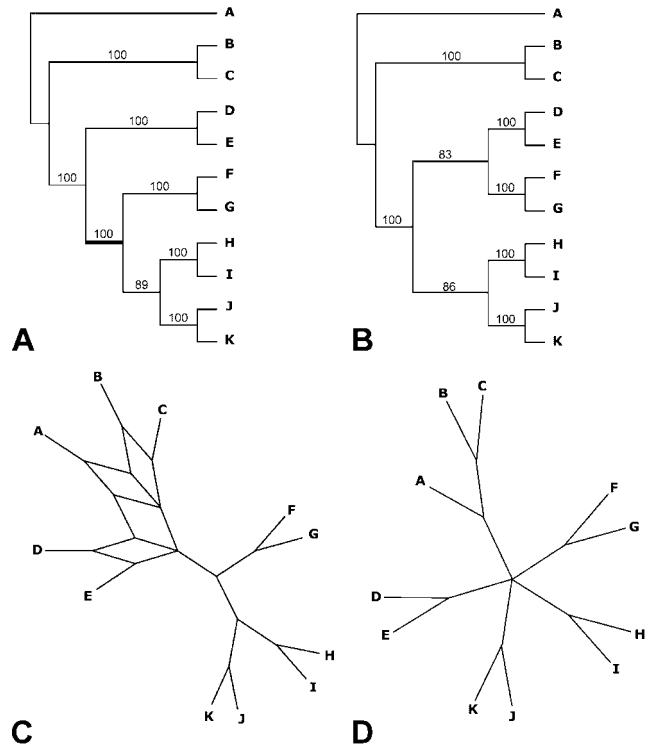


Fig. 3. Test of DRUIDS using simulated data generated using the changes in constraints depicted in Fig. 2. (A) Bootstrap tree (1000 replicates) obtained with all the simulated data; the thick branch is erroneous. (B) Bootstrap tree (1000 replicates) obtained with the simulated data minus regions detected by DRUIDS. Note that the inferred tree has the correct topology. (C,D) Split trees, respectively, for the whole data and the whole data minus DRUIDS regions (parsimony split and equal edges). Note that a split tree can appear unresolved while the phylogenetic inference shows resolution. It is due to the fact that split decomposition describes the conflicting signal rather than the sequence relationships.

results were obtained with maximum likelihood (results not shown).

Can DRUIDS identify regions of nonstationarity in real data sets?

Similar, but not identical, patterns of nonstationarity were observed for each of the four cytochrome *b* data sets. This suggests that there are constraints that are characteristic of the cytochrome *b* molecule but that these patterns may vary somewhat across taxa. The regions that were nonstationary for both hydrophobicity and volume were located in the middle of the same transmembrane helices (Fig. 4). Further analysis

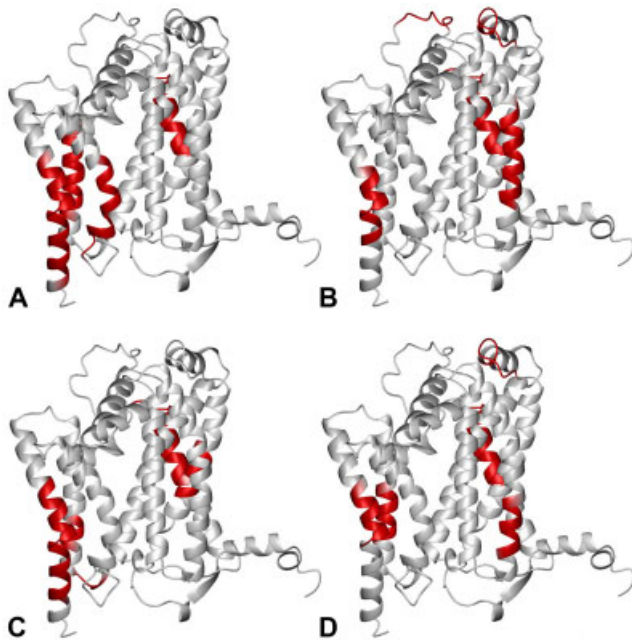


Fig. 4. Regions with deviation of stationarity (red) plotted on cytochrome *b* (bovine) (Xia et al., '97) for four independent data sets (window size = 21): (A) aves; (B) carnivores; (C) primates; and (D) teleosts. Graphics made with MolMol (Koradi et al., '96).

and interpretations of these patterns are presented elsewhere (Fedrigo et al., in preparation).

Can DRUIDS be used to improve phylogenetic accuracy?

Of the three original mitochondrial data sets only the *coI* data yielded a tree that was consistent with the “true” tree (Fig. 5). Chondrichthyans were paraphyletic for the *cytb* data set or were placed as the sister group to bony fishes in the *nd2* and combined data sets with 89% bootstrap support. Deleting DFS regions (40% of the data set) rendered all the data sets to be congruent with the expected tree, except for *nd2* in which several erroneously placed branches persisted albeit with weak bootstrap support. The combined data set minus DFS regions yielded the “true” tree with relatively high bootstrap support (Fig. 6). Application of DRUIDS to the *cytb* data set had broadly similar effects to those seen for the simulated data. Removal of misleading regions resulted in a loss in resolution but also less conflict and fewer erroneous relationships. We obtained consistent results using a maximum likelihood approach (results not shown).

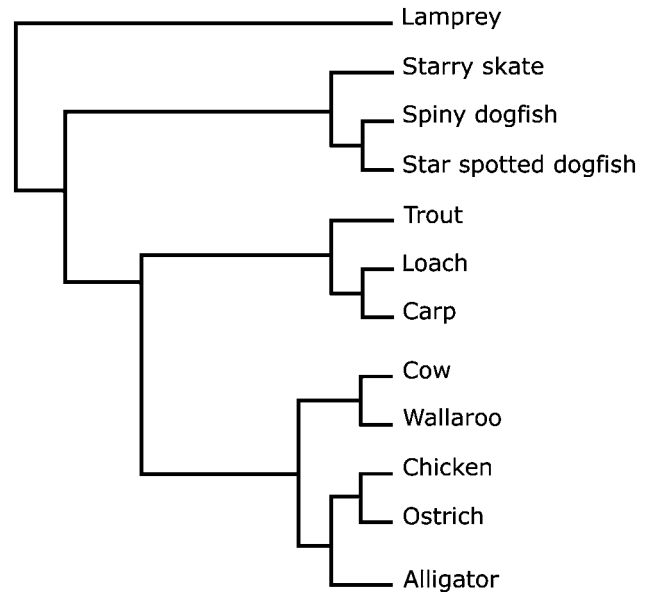


Fig. 5. “True tree” for the vertebrate data set.

Nonstationary processes can compromise the accuracy of phylogenetic estimation because most phylogenetic inference models assume stationarity (Saccone et al., '89; Lockhart et al., '92). However, there are cases where when nonstationarity can appear to strengthen the phylogenetic signal, such as in those cases where there has been a lineage-specific change in G+C bias. All the descendant taxa will share the same bias and have a stronger tendency than would otherwise be the case to “clump together” in a phylogenetic analysis. While this may result in a “correct” inference, it remains a violation of the model used and represents a case of obtaining right answer for the wrong reasons (e.g., Swofford et al., 2001). The DRUIDS algorithm detects nonstationary regions but does not differentiate whether the regions are phylogenetically misleading. In any case, nonstationary data is not desirable for phylogenetic reconstruction.

Deleting third codon position sites versus deleting DFS regions

Deleting the third position did not help to recover the “true” tree. On the contrary, bootstrap support for a (chondrichthyan–teleost) clade actually increased when third positions were removed. Split decomposition (Fig. 7) revealed that deleting the third position created proportionately more conflict than was otherwise detected. Figure 8A shows the relationship inferred from an

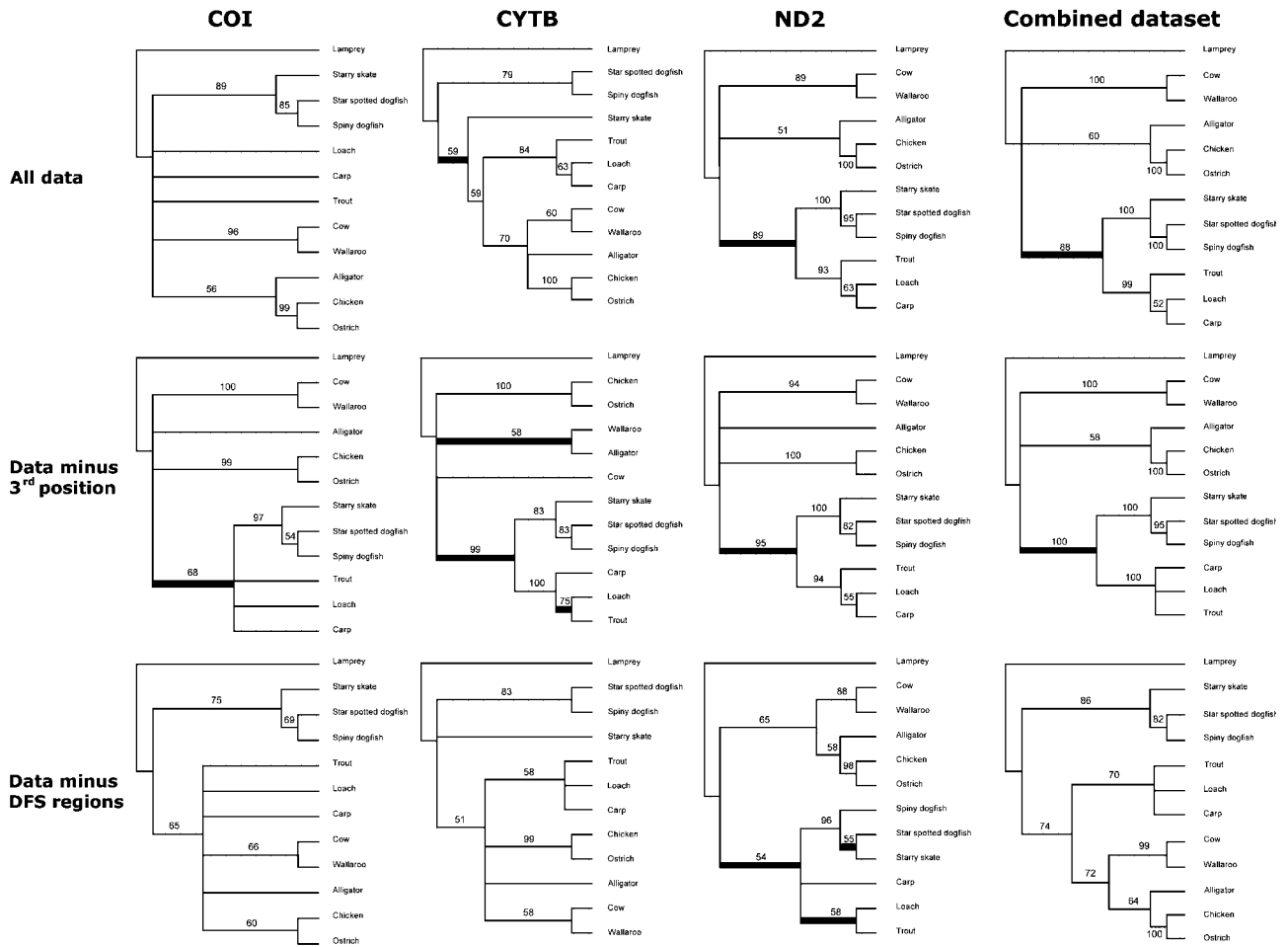


Fig. 6. Inference of MP trees with bootstrap values. The thick branches are erroneous.

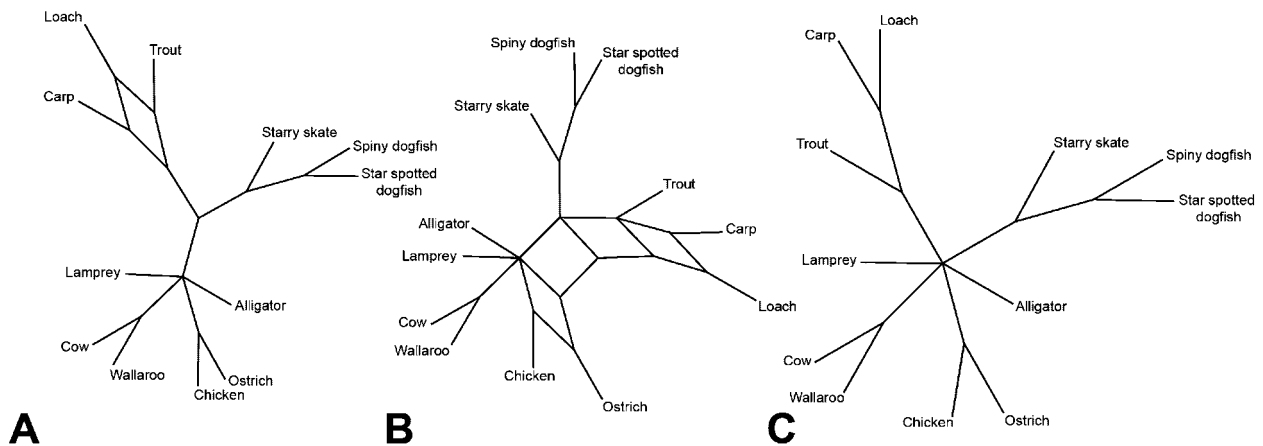


Fig. 7. Split trees (parsimony split and equal edges) for the combined data set (coI + cytb + nd2) for (A) all data; (B) all data minus 3rd position; (C) all data minus DFS regions.

analysis of third positions alone. The long terminal branches, short internodes, and lack of support for deep nodes are characteristic of fast-

evolving characters. The inferred topology was consistent with the “true” tree. Deleting third positions appeared to ameliorate base composition

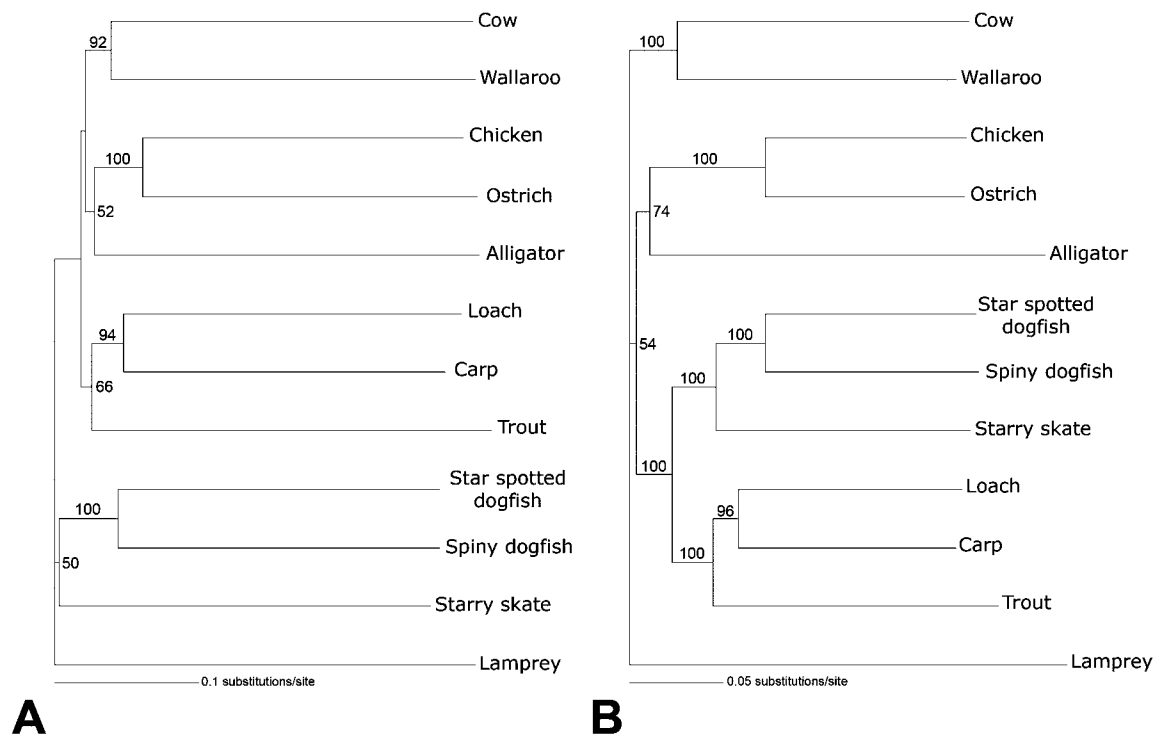


Fig. 8. Phylograms with bootstrap values (1000 replicates and uncorrected p distance) for purportedly misleading sites (i.e., those excluded from the analysis): (A) 3rd position only for the three genes; (B) DFS regions only for the three genes.

TABLE 2. Data set size, base composition bias (P value), and retention index (RI)

	coI	cytb	nd2	Combined data set
All data	1530 bp $P=0.00000001$ RI=0.2970	1128 bp $P=0.00000000$ RI=0.3329	939 bp $P=0.00000000$ RI=0.3722	3597 bp $P=0.00000000$ RI=0.3293
All data minus 3 rd position	1020 bp $P=0.99997413$ RI=0.4795	752 bp $P=0.71167085$ RI=0.4904	626 bp $P=0.01826302$ RI=0.4736	2398 bp $P=0.01678712$ RI=0.4741
All data minus DFS regions	1018 bp $P=0.00157052$ RI=0.2787	645 bp $P=0.00317791$ RI=0.3282	468 bp $P=0.03012778$ RI=0.3689	2131 bp $P=0.00000000$ RI=0.3090

bias (Table 2) and reduce the homoplastic effects associated with high rates of change. We refer to this type of homoplasy as Type I homoplasy. In general, Type I homoplasy is broadly dispersed in the tree, induced by an overall increase in evolutionary rate, and, if the taxon sampling is appropriate, does not lead to systematic error. Table 2 shows that the exclusion of this set of characters (3rd positions) improved the retention index, RI (Farris, '89). However, the inferred tree was still not consistent with

the true tree (Fig. 6), arguably because the removal of 3rd position sites did not account for undetected molecular convergences. When we examined the hierarchical signal contained within the nonstationary regions of the alignment, we obtained a tree with strong bootstrap support for all of the controversial nodes (Fig. 8B). This partition had a strong misleading signal that we suspect results from convergent constraints. We refer to this type of homoplasy as Type II homoplasy.

Type I homoplasy emphasizes chance convergences associated with overall fast rates of evolution or long time spans and can be thought of as white noise. If the “true” phylogenetic signal is strong enough, it will rise above this noise as championed by the “total evidence” school. The second type of conflicting signal, Type II homoplasy, is more problematic because it involves convergence due to independent changes of structural or functional constraint and is less readily identified as it need not be associated with high rates of change. Type II homoplasy is rarely randomly distributed over a tree and should be viewed as a directional homoplastic force with a tendency to attract unrelated lineages together. If it is stronger than the background noise and the “true” phylogenetic signal, it can be misleading. Long branch attraction (Felsenstein, '78), due to the independent accumulation of similar character states on two unrelated branches, can arise as a consequence of convergent constraints (causing type II homoplasy). The effect is exacerbated when changes in constraint are associated with elevated substitution rates. At its most fundamental, Type II homoplasy arises as a consequence of model misspecification. However, because it arises from nonstationarity of the substitution process, it generally cannot be fixed with a more parameter-rich stationary model. Rather, its appropriate amelioration requires that different models be assigned to different sequence regions in different parts of a tree. Deciding which models should be applied when and where is generally not immediately obvious from the aligned input data. However, we believe this will be a fruitful avenue for future research.

Are there alternatives to character deletion?

Some sites can be homoplastic for certain clades and synapomorphic for others. Deleting characters can be a risky strategy because, in removing homoplasy associated with one part of a tree, one might unwittingly be removing phylogenetic information in different part of the tree. The simulation study (Figs. 2 and 3) revealed that some support is lost after removal of the areas detected by DRUIDS. Patch II contains numerous synapomorphies for the clade ((H, I), (J, K)), even though it contains homoplastic character states for the taxa F and G. The effect of deleting regions can be even worse if the deleted regions do not represent the major source of bias in the data set

[e.g., when a data set is plagued with several long branches (Felsenstein, '78), or a complex distribution of among site rate variation]. In such situations, the decrease in synapomorphic support can pave the way for other biases to mislead phylogenetic reconstruction. What is needed is a method that down weights the phylogenetically misleading influences of a character in one part of a tree while leaving its phylogenetically informativeness intact in the remainder of the tree. This is best achieved through an explicit model of molecular evolutionary dynamics under a likelihood framework. This is an area of on going research.

Caveats

Sliding windows approaches are useful for detecting patterns of nonstationarity over multiple alignments of DNA sequences. In this paper, we show that the approach can be extended to identify shifts in biochemical constraints by considering nonstationarity at levels above that of nucleotide variation. The most appropriate window size is that commensurate with the scale of the biological constraint. Unfortunately, this is rarely apparent at the outset of any study. Thus we recommend setting a lower bound on window size equivalent to one turn of an alpha helix (four amino acids or 12 nucleotides) and an upper bound of two turns of an alpha helix (seven amino acids or 21 nucleotides). While we believe the approach represents an improvement over prior approaches, we caution that moving-window approaches are prone to “dilute” patterns of nonstationarity that are not distributed contiguously along a sequence. For example, if one face of a helix is exposed to solvent while the other is buried, the amino acids on the exposed surface would be constrained in a different way than those that are buried. Constraints associated with such a scenario would not be distributed contiguously along a sequence but would be distributed with a periodicity consistent with the turns of the helix. The averaging approaches intrinsic to moving-window analyses might cause such patterns to be missed. However, it is possible to detect these patterns by comparing the outcome of various window sizes. Indeed, large window sizes tend to smooth out variations in stationarity, whereas small window sizes will show more fine-grained and patchy distributions. Thus, a noncontiguously distributed constraint will be characterized by a marked difference in F -test score profiles between analysis of various window sizes.

We plan to develop approaches that consider windows that are composed not only of sequential neighbors but also of residues that are spatially adjacent on the protein structure. This would better accommodate patterns of constraint change that are spatially localized on the protein structure but sequentially distant in the sequence. While the DRUIDS approach to multiple alignment exploration is free from any ad hoc phylogenetic hypothesis, it does not identify all signals that are phylogenetically misleading. For example, DRUIDS does not account for problems associated with taxon sampling (Kim, '96; Graybeal, '98; Hillis, '98; Poe, '98). However, we believe that it does offer new opportunities for exploring the effects of constraint change, codon usage, and G+C bias on the phylogenetic signal.

ACKNOWLEDGMENTS

We thank Günter Wagner and two anonymous reviewers for helpful comments on the manuscript. We thank Andrés López for his input concerning the codon bias features of the software, all the people who tested different versions of DRUIDS, and Anna Keyte for editing various versions of the manuscript.

LITERATURE CITED

- Farris JS. 1989. The retention index and the rescaled consistency index. *Cladistics* 5:417–419.
- Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27:783–791.
- Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725–736.
- Grassly NC, Rambaut A. 1997. A likelihood method for the detection of selection and recombination using sequence data. *Mol Biol Evol* 14:239–247.
- Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* 47:9–17.
- Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. *Bioinformatics* 18:500–501.
- Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst Biol* 47:3–8.
- Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14:68–73.
- Kim J. 1996. General inconsistencies conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Syst Biol* 45:363–374.
- Koradi R, Billeter M, Wüthrich K. 1996. MolMol: a program for display and analysis of macromolecular structures. *J Mol Graph* 14:51–55.
- Kyte J, Doolittle RF. 1982. A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 157:105–132.
- Lanavé C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. *J Mol Evol* 20:86–93.
- Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. *J Mol Evol* 34:153–162.
- Long M, Gillespie JH. 1991. Codon usage divergence of homologous vertebrate genes and codon usage clock. *J Mol Evol* 32:6–15.
- Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11:715–724.
- Naylor GJP, Gerstein, M. 2000. Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. *J Mol Evol* 51:223–233.
- Page RDM, Cotton J. 2002. Vertebrate phylogenomics: reconciled trees and gene duplications. *Pac Symp Biocomp* 536–547.
- Poe S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. *Syst Biol* 47:18–31.
- Rasmussen A-S, Arnason U. 1999. Molecular studies suggest that cartilaginous fishes have a terminal position in the piscine tree. *Proc Natl Acad Sci USA* 96:2177–2182.
- Saccone C, Pesole G, Preparata G. 1989. DNA microenvironments and the molecular clock. *Mol Evol* 29:407–411.
- Storm CE, Sonnhammer EL. 2003. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res* 13:2353–2362.
- Swofford DL. 1999. PAUP*: phylogenetic analysis using parsimony, ver. 4.02ba. Sunderland, MA: Sinauer Associates.
- Swofford DL, Olsen GL, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Morowitz C, Mable BK, editors. *Molecular systematics*, 2nd edition. Sunderland, MA: Sinauer Associates. p 407–514.
- Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol* 50:526–539.
- Takezaki N, Goboiori T. 1999. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. *Mol Biol Evol* 16:590–601.
- Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* 87:23–29.
- Xia D, Yu C-A, Kim H, Xia J-Z, Kachurin AM, Zhang L, Yu L, Deisenhofer J. 1997. Crystal structure of the cytochrome bc1 complex from bovine heart mitochondria. *Science* 277:60–66.
- Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogenous selection pressure at amino acid sites. *Genetics* 155:431–449.
- Yokoyama S, Radlwimmer FB. 1998. The "five-sites" rule and the evolution of red and green color vision of mammals. *Mol Biol Evol* 15:560–567.
- Zamyatin AA. 1972. Protein volume in solution. *Prog Biophys Mol Biol* 24:107–123.