# Accuracy through Subsampling of Protein EvolutioN: Analyzing and reconstructing protein divergence using an ensemble approach

Roman Sloutsky[1,2], Kristen M. Naegle[1,*]

**1 Biomedical Engineering Department and the Center for Biological Systems Engineering, Washington University in St. Louis**
**2 Division of Biology and Biomedical Sciences, Washington University in St. Louis**

*** knaegle@wustl.edu**

## Abstract

Mapping the history of gene duplications which gave rise to a protein family encoded in a genome (a set of paralogs) can be critical to understanding how those proteins function in their host cells today. However, since each member of a family is recapitulated in the genomes of related species (a set of orthologs), selection of sequences to be included in the history reconstruction is non-trivial. Reconstruction is extremely sensitive to the choice of sequences, which is deeply problematic given no mechanism exists for assessing the accuracy of individual reconstructions. Here, we capitalize on the variability of phylogenetic tree reconstruction to selected input sequences, by subsampling from the available ortholog sequences of a protein family to create an ensemble of trees, which explores the space of plausible tree topologies. We hypothesize that the most consistent topological features across an ensemble are more likely to be true and propose a tree reconstruction algorithm (ASPEN) based on this hypothesis. We simulate 600 protein families over known phylogenies, with varying branch lengths, and compare the accuracy of ASPEN reconstructions to those of traditional phylogeny inference methods. We find that ASPEN trees are more accurate than trees reconstructed traditionally. Additionally, we develop an observable metric calculated form subsampling, reconstruction Precision, for assessing the likely accuracy of a traditional, single-alignment all-sequence reconstruction of the divergence history for a set of paralogs. Together these findings suggest that an ensemble of imperfect reconstructions can provide more accurate insight than any individual reconstruction.

## Introduction

Protein families grow in size and diversity through duplication of genes encoding existing family members followed by functional divergence of the duplicates [1, 2]. Immediately following a gene duplication event the affected genome contains two identical copies of the duplicated gene. Because the genes are redundant, relaxed purifying selection allows mutations to accumulate rapidly. Since the added energy cost of expressing identical products from redundant loci confers a selective disadvantage, mutations resulting in loss of functionality by one of the copies are typically favored by selection. However, the rapid accumulation of mutations can also result in partial or complete functional divergence between the two copies. This may create a selective advantage due to increased functional repertoire through neo-functionalization, greater efficiency and control through sub-functionalization, and possibly resistance to deleterious mutations through vestigial functional overlap (functional moonlighting) [3–5], leading to retention of both diverged copies (paralogs). After subsequent speciation events give rise to diverged genomes (species), each of those genomes contains a gene descended through speciations from each paralog in the ancestral genome (Figure 1). These genes are orthologs characterized by a "same gene, different genome" relationship. Ortholog sets are related to each other as paralogs, since their respective Most Recent Common Ancestors (MRCAs) were the original paralogs in the ancestral genome. The genome of each species encodes a paralog gene belonging to each ortholog set.

Reconstructing the divergence history (topology) of a protein or protein domain family is crucial to understanding the proteins' (protein domains') function(s) and evolution. In addition to facilitating powerful *in silico* analyses [6–13], reconstructions of paralog divergence guide experimental design and data interpretation [14–19]. Accordingly,

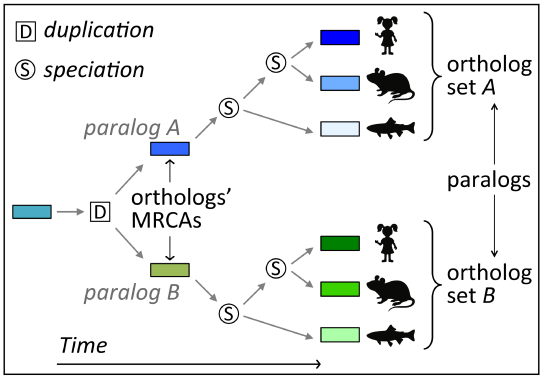


**Figure 1. A hypothetical protein divergence history.** Two paralogs emerge after a duplication event and are passed along through subsequent speciation events. If no additional duplication events occurred, paralogs *A* and *B* existed at one time as Most Recent Common Ancestors (MRCAs) of two ortholog sets and exist today in the genomes of species emerged through the series of speciations. Each ortholog set can be thought of as representing its MRCA's paralog.

divergence reconstructions for well-studied protein domain families [20–22] have been relied upon extensively by the scientific community. Because such reconstructions are created from single sequence alignments, they ignore the great deal of uncertainty in topology reconstruction under equally valid alignment representations of input sequence data.

Divergence topology reconstruction is extremely sensitive to the input alignment. For example, the same sequences aligned by different algorithms [23–27] or using different guide trees [28] yield different topology reconstructions. So does reversing input sequences prior to alignment [29, 30], or removing less than 0.1% of columns from an alignment containing over 600,000 columns [31]. For paralog divergence topologies, another source of uncertainty likely to influence reconstruction is the set of orthologs selected to represent each paralog. Because duplications usually predate numerous speciation events, they tend to correspond to deep internal nodes – nodes with many descendant leaves – in full phylogenies of protein families. MRCAs of ortholgs descend from duplications (Figure 1), meaning every ortholog descended from each MRCA is also descended from the duplication. Deep internal nodes tend to be most sensitive to perturbations of the input alignment [32]. Unfortunately, since the true history of protein divergence is hidden from us in time, we have no way of knowing which divergence topologies are more accurate, given the equal validity of input alignments.

Although traditional tree reconstruction produces phylogenies – topologies parametrized with branch lengths reflecting extent of divergence – we disregard the branch lengths here to focus on the topologies alone. In traditional inference topologies and branch lengths are inferred jointly, alternating between topology modifications and branch length optimization in the case of statistical (Maximum Likelihood and Bayesian) methods. Because the likelihood function is evaluated many times for each proposed topology, and topology space is almost unfathomably large, statistical methods can suffer extremely long run times on large sequence collections. However, if accurate candidate topologies can be identified by other means, the computational cost of optimizing branch lengths for individual topologies is nearly trivial, while optimization for multiple topologies is embarrassingly parallel. Our approach permits separating topology reconstruction from branch length optimization.

Furthermore, we focus on reconstructing only the hardest topology nodes – the deep internal nodes corresponding to protein or domain paralog divergence. We treat MRCAs of ortholog sets as leaves in our reconstructions and disregard ortholog divergence, which overwhelmingly recapitulates the species tree. Species divergence is reconstructed more accurately by other approaches [31, 32]. Instead, we capitalize on the variance in reconstructed topologies under changes in ortholog representation of paralogs to separate topological features we believe to be supported by phylogenetic signal from spurious ones we believe to result from noise. We hypothesize that features observed more frequently under ortholog resampling are more likely to reflect signal and, therefore, be more accurate, than less frequently observed ones. We explore the relationship between accuracy and variability in reconstructing paralog divergence topologies and propose a metric for assessing the likely accuracy of a single-alignment reconstruction for a given protein family. We then present ASPEN, a topology reconstruction algorithm that integrates over the uncertainty of single alignment reconstructions to build and rank trees according to observations across reconstructions from many equally valid alignments. ASPEN produces more accurate topologies than

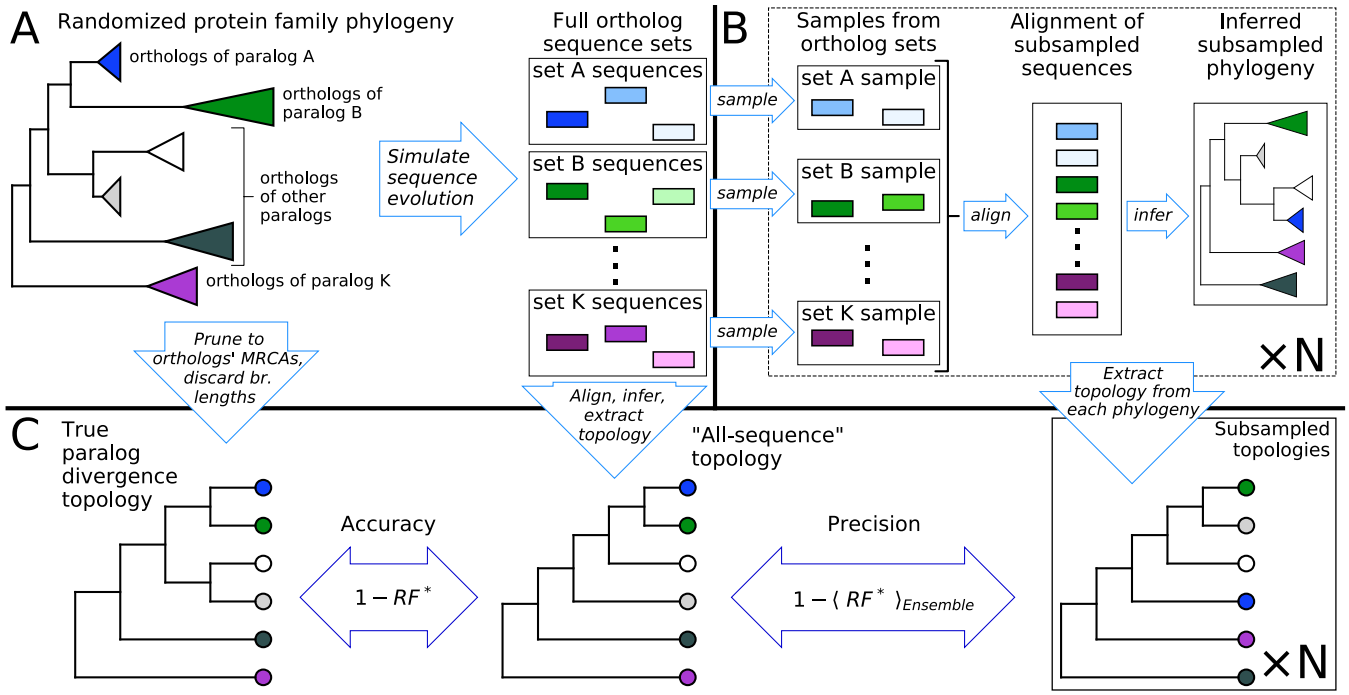


**Figure 2. Analysis framework for comparing reconstruction Accuracy and Precision.** (**A**) Sequence evolution was simulated over synthetic phylogenies. Synthetic phylogenies were pruned to MRCAs of ortholog sets and branch lengths were discarded to obtain true paralog divergence topologies. Simulated sequences were aligned, phylogenies were inferred from those alignments, and "all-sequence" reconstructions of paralog divergence topologies were extracted. (**B**) Sequences were repeatedly sampled from each ortholog set in a family and phylogeny inference and topology extraction were done to produce a "subsampled topology". Repeating this N times yields an ensemble of topologies. (textbfC) We define Accuracy as the similarity between the all-sequence reconstruction and Precision as the comparison between subsampled topologies and the all-sequence topology.

traditional reconstructions from single, all-sequence alignments.

## Experimental framework for reconstruction analysis

We generated test sequence data by simulating evolution of protein families instead of using natural protein sequences for two previously noted reasons [25]. First, simulating evolution over known phylogenies allowed us make a quantitative assessment of reconstruction accuracy compared to the "true" divergence topology. Second, it allowed us to explore a range of divergence conditions by systematically varying branch lengths of input phylogenies, while controlling for other factors such as overall sequence length and the distribution of secondary structure elements and disordered loops. Assembling a comparable biological data set would have been impossible.

We simulated families containing 15 paralogs, each represented by 66 orthologs. In order to make the assessment statistically robust, we generated 600 families across a range of post-duplication branch lengths. An alignment of human tyrosine kinase domains (median length 269 a.a.) was used as template for all simulations (see *Methods* for simulation details). We then used all combinations of three multiple sequence alignment algorithms (MAFFT's L-INS-i protocol [33], ClustalOmega [34], and Muscle [35]) and two phylogeny inference algorithms (FastTree2 [36] and RAxML [37]) to reconstruct phylogenies for the 600 simulated families. We compared the reconstructed paralog divergence topologies, excluding speciation nodes by pruning orthologs' MRCAs to leaves, to the true divergence topology over which evolution was simulated (Figure 2A). We quantified topology differences with the Robinson-Foulds symmetric distance metric [38], modified to account for the occasionally non-monophyletic reconstruction of ortholog sets ($RF^*$, *Methods*). For convenience we define the accuracy of a reconstruction as $1 - RF^*$ distance between reconstructed and true paralog divergence topologies. Consistent with earlier studies [24–27, 39–41], choice of alignment algorithm substantially affected accuracy, with L-INS-i alignments

producing most accurate reconstructions, while FastTree2 and RAxML performed very similarly across all alignments (Figure *S1*). Based on these results, we selected the combination of L-INS-i and FastTree2 for all remaining analysis.

## Subsampling reveals an observable measure of accuracy

Given the known sensitivity of reconstruction to input alignment, we explored reconstruction variance resulting from differences in ortholog representation of paralogs using the framework outlined in Figure 2. We gathered the sets of ortholog sequences representing each paralog in a simulated family (Fig. 2A) and performed a resampling experiment (Fig. 2B): 50 times we randomly sampled 60 of 66 sequences (91%) from each ortholog set and performed traditional reconstruction, using L-INS-i and FastTree2 with each collection of subsampled input sequences. We retained a large fraction of sequences to minimize both the input variation and the loss of phylogenetic signal. To quantify reconstruction uncertainty, we measured the similarity (1− the average of $RF^*$ ) between topologies reconstructed from most of the sequences to the "all-sequence" topology (Fig. 2C). Since this quantity is a measure of how close the estimates are to each other, we refer to it as Precision.

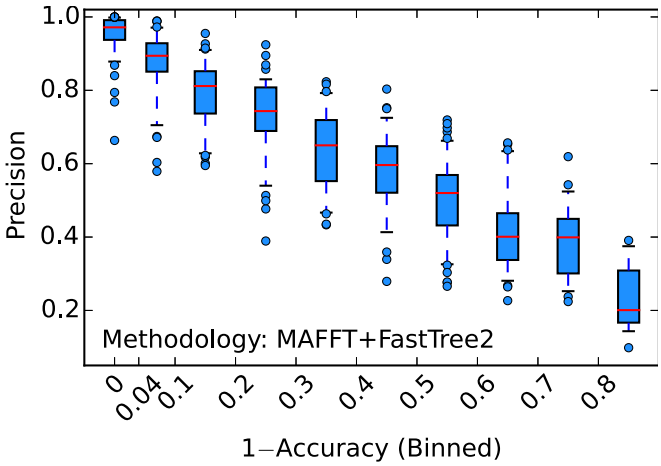


**Figure 3. Precision vs Accuracy of reconstruction.** Reconstruction Precision plotted vs 1−Accuracy of all-sequence reconstruction for each simulated protein family. 1−Accuracy used on x-axis to make families with most accurate reconstructions appear on the left and those with least accurate on the right. Families were binned by 1−Accuracy. Tick marks on x-axis indicate bin boundaries.

Figure 3 demonstrates the striking relationship between accuracy of the all-sequence reconstruction (Accuracy) and Precision of reconstruction for families across a range of post-duplication branch lengths. Due to their strong correlation we use Precision, an observable quantity for natural protein families, as a measure of a family's reconstruction Accuracy (unknowable for natural proteins) and, by proxy, the overall "complexity" of reconstruction for that family. Importantly, this also suggests that our 600 synthetic protein families span a range of complexities, allowing us to observe the performance of reconstruction as a function of complexity, via its proxy – Precision.

## Using variability to distinguish phylogenetic signal from noise

Although we observed high reconstruction Precision for many families, only four of 600 families had identical paralog divergence topologies reconstructed from every subsampled alignment (Precision=1). Even among families with the highest Precision, and under dense subsampling, reconstruction variability was pervasive. On the other hand, Salichos and Rokas [32] argued that pairwise $RF$ distances smaller than 1 (the average $RF$ distance among randomly generated topologies) indicates consistent phylogenetic signal among the topologies being compared. Most of our 600 families had Precision ($1-\langle RF^* \rangle$) significantly greater than 0, but less than 1. Thus we sought to go a step further and test our central hypothesis: not only does intermediate Precision indicate consistent signal, but more frequently recapitulated features are more likely to be accurate, and this fact can be used to reconstruct more accurate topologies. In order to test this we first needed a way to extract frequently recapitulated features, and then a way to identify topologies most consistent with those features. Next we describe our method, ASPEN, which accomplishes both tasks.

# Reconstructing topologies from ensemble sampling

We created a method we call ASPEN, for Accuracy through Subsampling of Protein EvolutioN, to construct and score topologies according to their consistency with topological features frequently represented in an ensemble of subsampled reconstructions (Fig. 2B). It relies on two key innovations: 1) extraction of topological features from an ensemble as frequencies of path lengths between leaves, and 2) an algorithm to construct and score topologies according to their consistency with observed path length frequencies.

## Transforming topology sets into path length distributions

ASPEN's foundation is the equivalent representation of a topology (an acyclic, bifurcating graph) as a matrix of path lengths between leaves in terms of the number of internal nodes encountered along a path. First we demonstrate equivalence of graph and matrix representations by presenting a simple algorithm for interconverting between the two (Figure 4). Then we discuss how ensembles of topologies are transformed into path length frequency distributions.

### Transforming a topology graph into a path length matrix

A topology can be equivalently represented as a matrix of leaf-to-leaf path lengths in terms of internal nodes encountered along the path. Transformation of a topology into its path lengths matrix representation is trivially accomplished by counting internal nodes along each path between pairs of leaves (Figure 4A).

### Transforming a path length matrix into a topology graph

The reverse transformation can be accomplished using a simple bottom-up construction procedure (Figure 4B). Internal nodes are constructed by joining pairs of leaves and/or previously constructed internal nodes to recapitulate observed leaf-to-leaf path lengths. This bottom-up construction ("outside-in" for unrooted topologies) continues until all leaf nodes are connected by a single graph. Note that it is possible to encounter path lengths during list traversal which, at that state of construction, cannot be accommodated by constructing an internal node. For example, if the order of paths $(A \leftrightarrow E, 3)$ and $(A \leftrightarrow F, 3)$ in the list in Figure 4B were reversed and path $(A \leftrightarrow F, 3)$ was encountered first, it could not be accommodated because internal node $\{\{A, B\}, \{\{C, D\}, E\}\}$ would not yet be available to join to leaf F. Such path lengths are skipped and then revisited on the subsequent traversal of the list. Traversal is repeated as necessary until construction is completed. Because all path lengths are derived from a single topology, they are guaranteed to be consistent, making the construction unambiguous.

### Generating path length frequency distributions

We take advantage of the alternate matrix representation to capture the individual variation of each leaf-to-leaf path length across an ensemble of topologies. Each topology is transformed into a path lengths matrix. Then path lengths for each pair of leaves are aggregated into a path length distribution for that pair (Figure 4C). Although ortholog sets overwhelmingly group into monophyletic subtrees across ensemble topologies (their MRCAs have no descendant leaves besides themselves), occasionally reconstructions do yield non-monophyletic ortholog sets. Because this violates an underlying assumption of the reconstruction, as well as the true topology of each synthetic protein family, we preclude paths compromised by this incorrect reconstruction from contributing to path length distributions: the length of any leaf-to-leaf path that contains a compromised internal node is not included in the distribution for that leaf pair.

## Path length frequencies guide topology reconstruction

### A score reflecting consistency with extracted features

ASPEN uses a quantitative metric for measuring the consistency of a proposed topology with observations from an ensemble of topologies. The score assigned to a topology is expressed in terms of log frequencies of leaf-to-leaf path lengths, $\log(f^L_{pair})$ where $L$ is the length of path between leaves in *pair*, incorporated into the topology:

$$score = \sum_{\substack{leaf \\ pairs}} \log(f^L_{pair})$$

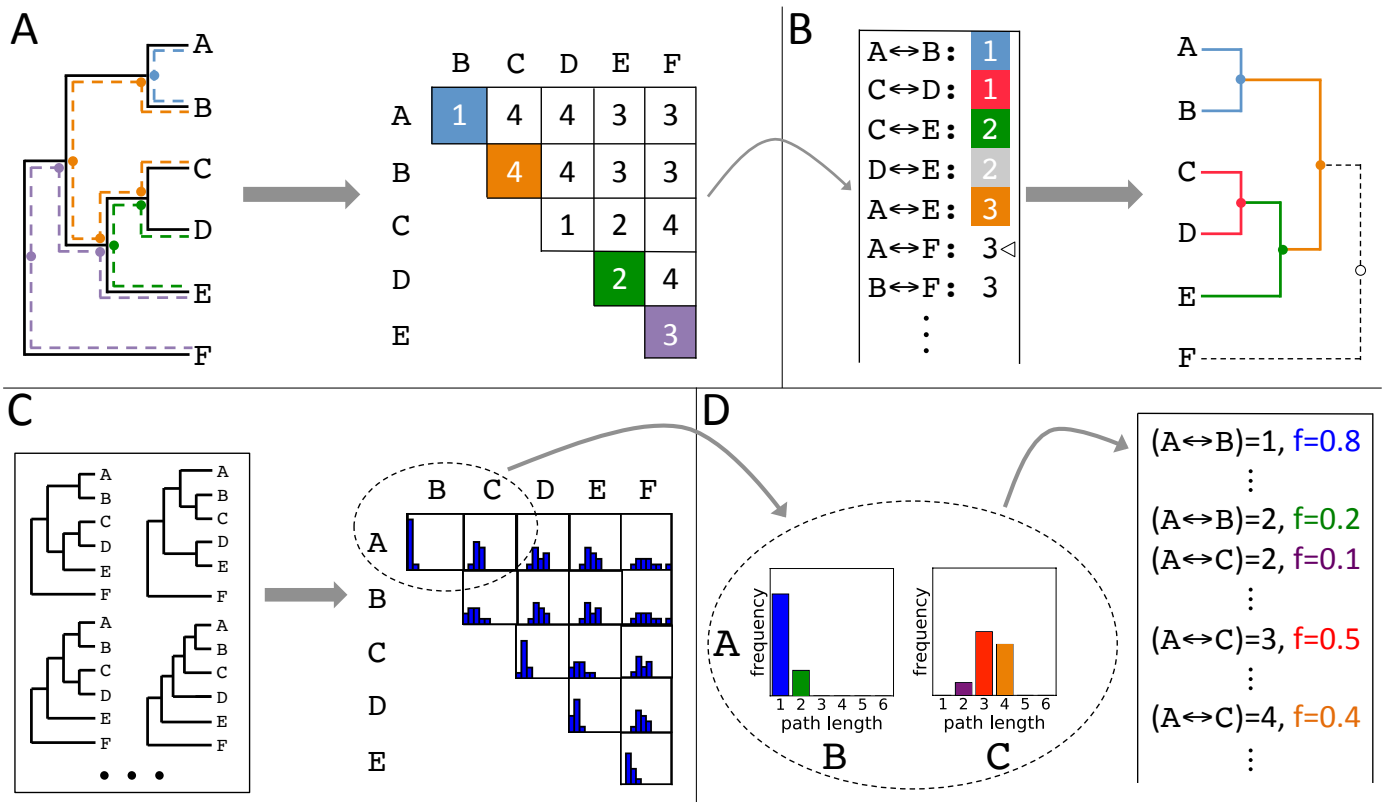


**Figure 4. Aggregating topological features across an ensemble of topologies using the path lengths matrix representation** (**A**) Decomposition of a topology into a matrix of leaf-to-leaf path lengths. Sample paths (A↔B,1), blue, (D↔E,2), green, (E↔F,3), violet, and (B↔C,4), orange, are highlighted. Dots indicate internal nodes along path. (**B**) Construction of a topology from a matrix of path lengths. First, the matrix is transformed into a sorted list of path lengths. Construction of internal nodes is triggered by path lengths encountered traversing the list: 1) Node $\{A, B\}$ joins leaves A and B and completes path $(A \leftrightarrow B, 1)$, blue. 2) Node $\{C, D\}$ joins leaves C and D and completes path $(C \leftrightarrow D, 1)$, pink. 3) Node $\{\{C, D\}, E\}$ joins leaf E to internal node $\{C, D\}$ and completes path $(C \leftrightarrow E, 2)$, green. Path $(D \leftrightarrow E, 2)$, grey, is completed by the same node and can be skipped during list traversal. 4) Node $\{\{A, B\}, \{\{C, D\}, E\}\}$ joins internal nodes $\{A, B\}$ and $\{\{C, D\}, E\}$ and completes path $(A \leftrightarrow E, 3)$, orange. Four paths of length 4 which appear further down the in the list are also completed by this node. Finally, 5) node $\{\{\{A, B\}, \{\{C, D\}, E\}\}, F\}$ joins leaf F to internal node $\{\{A, B\}, \{\{C, D\}, E\}\}$ and completes path $(A \leftrightarrow F, 3)$, dashed line. This completes the reconstruction, since all leaves are connected by the resulting topology. Path $(B \leftrightarrow F, 3)$ and all subsequent paths are already completed and can be ignored. (**C**) Each topology in the ensemble is decomposed into a matrix of leaf-to-leaf path lengths. Observed path lengths for each pair of leaves are aggregated into distributions. (**D**) Each distribution is then converted into a set of constraints on the length of the path between that pair of leaves. In the expanded section of the path lengths matrix, distributions of lengths for paths $(A \leftrightarrow B)$ and $(A \leftrightarrow C)$ are turned into constraints on the lengths of these paths by inserting each observed distance for each path, together with the frequency with which that distance was observed, into a list of path lengths. Vertical ellipses represent other paths of lengths 1, 2, 3, 4, etc. coming from elsewhere in the matrix.

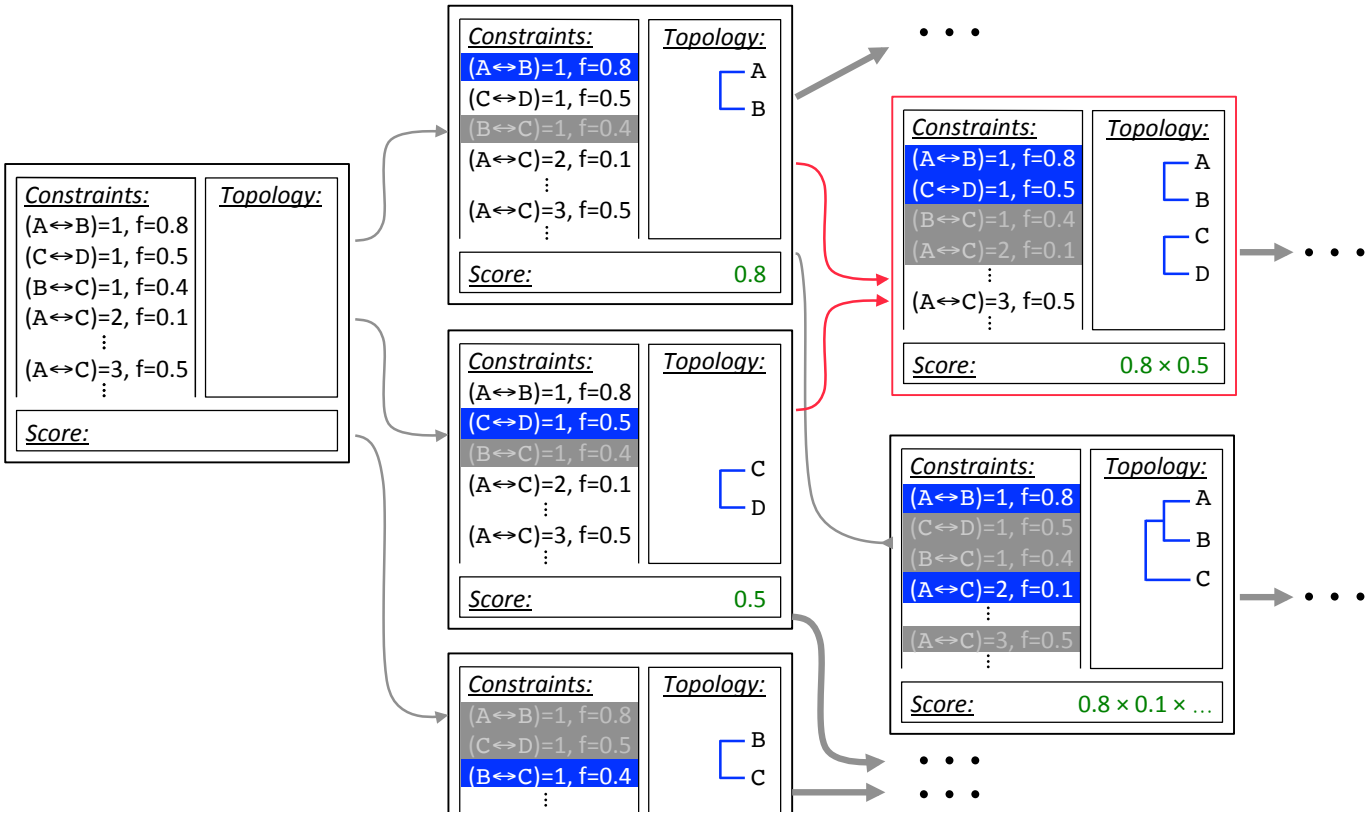


**Figure 5. Branching construction of topologies by incorporating path lengths observed in an ensemble.** Construction begins with the empty topology assembly on the left. Every possible extension is constructed in a copy of the initial assembly: Node $\{A, B\}$ completes path $(A \leftrightarrow B, 1)$, node $\{C, D\}$ completes path $(C \leftrightarrow D, 1)$, and node $\{B, C\}$ completes path $(B \leftrightarrow C, 1)$, branching the initial assembly into three new assemblies. Path lengths completed by the introduced node and path lengths incompatible with it are marked and not revisited. Nodes $\{A, B\}$ and $\{C, D\}$ preclude path $(B \leftrightarrow C, 1)$, while node $\{B, C\}$ precludes paths $(A \leftrightarrow B, 1)$ and $(C \leftrightarrow D, 1)$. Completed paths are shown in blue, precluded paths are greyed out in the corresponding assemblies. Intermediate topology scores are calculated according to the scoring function. On the next iteration construction paths for assemblies $\{A, B\}$ and $\{C, D\}$ collide, indicated in red. A single copy of the resulting assembly, $\{A, B\}, \{C, D\}$, is retained. Assembly $\{A, B\}$ is separately extended with node $\{\{A, B\}, C\}$. Additional construction paths, indicated by ellipses, are not shown.

This scoring function rewards incorporation of frequently observed path lengths and penalizes rarely observed path lengths.

### A branch-and-bound topology construction algorithm

Using the bottom-up procedure for constructing a topology graph from its path lengths matrix representation (Figure 4B), we developed an algorithm that uses a branch-and-bound strategy to construct the requested number of highest-scoring topologies according to the scoring function above. We describe the branching and bounding procedures in the next two sections.

**Branching** By analogy with the single-topology procedure in Figure 4B, construction of internal nodes is triggered by path length entries encountered during list traversal. However, this list contains every observed path length for every leaf pair, together with its frequency (Figure 4D). Unlike the single-topology case, list entries cannot be assumed to be consistent with each other. In fact, many combinations of path lengths on the list cannot be incorporated into one topology. For example, for hypothetical leaves $A$, $B$, and $C$, path lengths $(A \leftrightarrow B, 1)$ and $(B \leftrightarrow C, 1)$ are mutually exclusive because in a bifurcating topology $B$ can be one internal node removed from either

$A$ or $C$, but not both. In single topology reconstruction, if a path length could be completed by the introduction of an internal node, that node could be safely constructed because it was guaranteed to satisfy every other list entry. Since that guarantee no longer holds, multiple topologies are constructed simultaneously by allowing the construction path to branch (Figure 5).

"Assemblies" are used to track simultaneous reconstruction of multiple topologies. Each assembly holds a copy of the path length frequencies list, a partially constructed topology, and the current topology score according to the scoring function (discussed below in the section on bounding). Reconstruction proceeds in iterations, starting with a single empty assembly (Figure 5, left). On the first iteration, the entire list is traversed and *every* possible extension by introduction of a new node is created simultaneously in a copy of the original assembly (Figure 5, middle). In each new assembly, all path lengths completed by the new node and all path lengths incompatible with it are marked and not re-examined on subsequent iterations. Remaining path lengths are not completed by the new node, but remain compatible with it. On subsequent iterations the same procedure is repeated for all tracked assemblies.

In principle, branching and iteration alone yield every topology consistent with path lengths observed in the ensemble. In practice, this results in a combinatorial explosion which must be carefully managed to allow construction to proceed to completion. First, Figure 5 (right) demonstrates how branching to satisfy non-conflicting path lengths can lead to collisions between diverged construction paths on later iterations. This occurs because many topologies can be constructed by introducing internal nodes in multiple orders. Each branched path represents a particular order of internal node introduction. In a practical implementation collisions must be managed in order to prevent multiple reconstructions of the same topology by multiple paths – an enormous replication of effort.

Second, even if each distinct topology is constructed once, in most cases reconstructing every topology consistent with observations from the ensemble, no matter how infrequent, is neither practical nor useful. Bounding, described in the next section, guarantees reconstruction of only the requested number of top scoring topologies.

**Bounding** The score is used to rank completed topologies, where ranking is updated every time a new topology is finished. The number of top scoring topologies to reconstruct, $X$, is requested at the beginning of a reconstruction run (10,000 was used in ASPEN evaluation). Once the initial $X$ topologies are constructed, the $X$th topology score constitutes the bound. Partially constructed topologies are abandoned if no complete topology that can be derived from that construction state will score above the bound. We determine this by calculating the score for already-incorporated path lengths and projecting the best possible score for a complete topology by assuming the most frequent remaining path length will be incorporated for every unconnected leaf pair:

$$projected = \sum_{\substack{incorporated \\ paths}} \log(f^L_{path}) \quad + \sum_{\substack{remaining \\ paths}} \max(\log(f^L_{path}))$$

As more high-scoring topologies are constructed, the bounding criterion becomes more strict allowing both more and earlier abandonment of unproductive construction paths. The branch-and-bound strategy guarantees that the $X$ topologies remaining on the list at the end of a run are the $X$ highest scoring topologies according to the scoring function.

# Evaluation and Discussion of ASPEN reconstructions

To test our algorithm, for each protein family we generated ensembles of 1000 subsampled topologies with each ortholog set represented by 30 of 66 orthologs (≈45%). Then we used ASPEN to reconstruct 10,000 top scoring topologies for two-thirds of the families. Because accuracies of all reconstructions vary substantially across the range of reconstruction Precision, as does the relative accuracy of ASPEN-reconstructed topologies, the families were binned by their Precision for the purposes of this analysis. Next we examine the relationship between reconstruction Precision and the discriminatory power of the log-frequency function with respect to accuracy, and then compare ASPEN reconstructions with all-sequence reconstructions and discuss the implications of our observations.

## Log-frequency score is correlated with accuracy

To understand the relationship between the log-frequency score and the accuracy of reconstructed topologies, we plotted the ASPEN topology rank vs the bin-average accuracy of topologies (Figure 6B-G). Among higher-Precision

families (Figure 6B-D), top-ranked log-frequency scores are strongly correlated with accuracy, particularly for topologies ranked in the top ∼50, which indicates the independent scoring function based on observed frequencies across the ensemble are indicative of accuracy. The strength of correlation decreases as difficulty of reconstruction increases (lower Precision bins, Figure 6E-G), indicating less discriminatory power with respect to accuracy. Nevertheless, ASPEN's top-ranked topology is, on average, also its most accurate across all Precision bins.

## Top ASPEN topology beats all-sequence reconstructions

Next, we compared ASPEN's best topology to all-sequence single-alignment reconstructions (Figure 6A). Like all other methods, ASPEN's accuracy is a function of Precision, or difficulty of the reconstruction task. As discussed earlier, MAFFT L-INS-i alignments yielded the most accurate all-sequence reconstructions across all Precision bins, while FastTree2 and RAxML performed very similarly on all alignments. Both top-ranked ASPEN topologies and L-INS-i all-sequence reconstructions have nearly perfect accuracy on families in the highest-Precision bin – not particularly surprising, since subsampled topology ensembles for ASPEN reconstruction were generated using the combination of L-INS-i and FastTree2 (*Methods*). Much more intriguing is the fact that top-ranked ASPEN topologies are consistently more accurate than any all-sequence reconstruction across the remaining Precision bins. Moreover, although the accuracy of all reconstructions degrades with difficulty of the reconstruction task (lower Precision), ASPEN's accuracy degrades much more slowly. ASPEN's top topology provides the greatest accuracy improvement over single-topology reconstructions when reconstruction is the most difficult.

## ASPEN produces many accurate topologies at low Precision

To compare more ASPEN topologies with the most accurate all-sequence reconstructions, bin-average accuracies of L-INS-i / FastTree2 topologies are plotted alongside bin-average accuracies of top-300 ranked topologies (Figure 6B-G). Although the log-frequency score provides less discrimination with respect to accuracy, more of ASPEN's topologies outperform single-alignment reconstructions as Precision decreases and reconstruction becomes harder. In the two lowest-Precision bins (Figure 6E-G), all top-300 ASPEN topologies are more accurate than the best all-sequence reconstruction. Taken together, these observations suggest that ASPEN results should be considered differently for families with high and low reconstruction Precision.

## How to use ASPEN in different Precision conditions

### The top few topologies are best for high-Precision families

For families with high Precision, where one may reasonably expect to reconstruct an accurate topology, ASPEN's top, or top few topologies are likely more accurate than any single-alignment reconstruction. One or a few of these topologies can be confidently used for downstream applications. This result is far from trivial, given that ASPEN's subsampling approach scales far better with the overall number of input sequences than traditional statistical reconstruction methods. With the advent of affordable genome sequencing and the resulting explosion in the number of sequenced and annotated species' genomes [42, 43], all-sequence reconstruction of paralog divergence by statistical methods has become infeasible for many families with large numbers of orthologs. Therefore, subsampling large samples of orthologs to yield a Precision score can now be used to identify how likely the full sequence topology is to be accurate, determining if one is working in a high or low Precision/Accuracy regime.

### Diverse representation is critical at lower Precision

Accuracy of all reconstructions suffers for families with lower reconstruction Precision (greater difficulty for reconstruction). Even top-ranked ASPEN topologies cannot be expected to be completely accurate. In this Precision regime all of the top 300 ASPEN topologies, or more, can be considered comparably plausible models, given the sequence data. Since all of these topologies are very likely to be more accurate than any individual single-alignment reconstruction, under these conditions ASPEN topology reconstruction should be treated as a mechanism for sampling a large number of imperfect, but quite accurate topologies. As the true topology cannot be distinguished from other, fairly accurate topologies on the basis of such sequence data, any downstream analysis relying on a divergence topology should aim to integrate over this topological uncertainty.

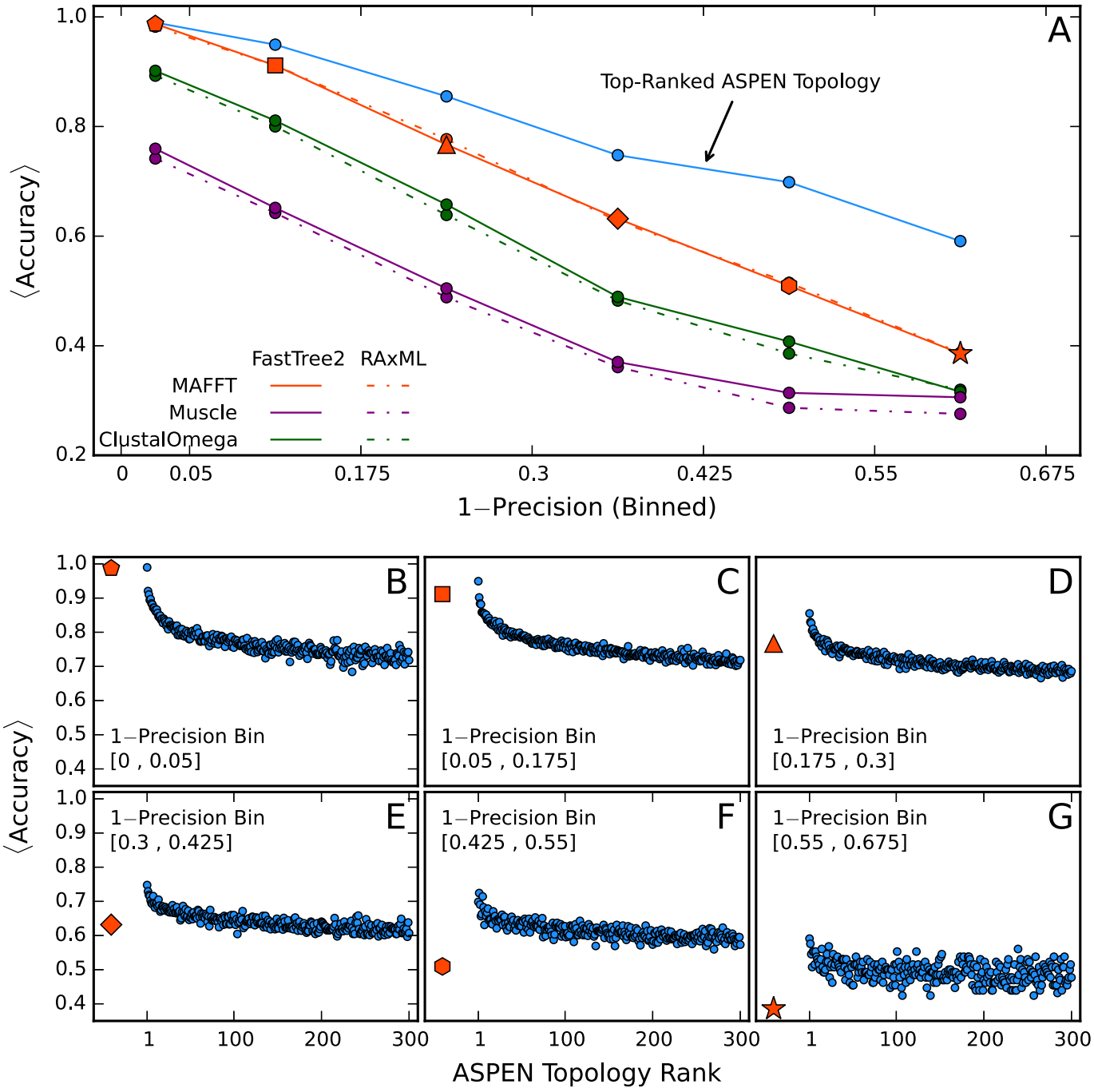


**Figure 6. Accuracy of topologies reconstructed by ASPEN.** (**A**) Accuracy, as a function of 1−Precision of a family's reconstruction, of the top-ranked ASPEN topology and all-sequence reconstructions. Families were binned according 1−Precision. Ticks on x-axis correspond to bin edges. Average accuracy of each type of reconstruction across families in the bin is plotted. For all-sequence reconstructions with MAFFT L-INS-i and FastTree2 (orange, solid line) a unique marker shape is used in each Precision bin. (**B**)-(**G**) For each Precision bin in (A), accuracy of ASPEN topologies ranked 1 through 300, averaged for each rank across all families in the bin, plotted as a function of rank. Average accuracy of the L-INS-i / FastTree2 all-sequence reconstruction is plotted for comparison on the left of each panel.

# Conclusion

Subsampling in the process of reconstruction proved to be extremely powerful – it identified two measures (Precision and Score based on observed frequencies) of something unknowable (Accuracy) and guided a reconstruction method that identifies much more accurate topologies than traditional approaches. That ASPEN reconstructions were more accurate than single-alignment reconstructions, is evidence that the central hypothesis of this work is supported – relationships found consistently amongst the variance produced by subsampling are more likely to be reflective of true protein divergence histories. We anticipate that, as a meta analysis approach to tree evaluation and reconstruction, ASPEN is likely to continue to boost the accuracy of individual approaches.

We also conclude from this study that it is worth revisiting the reconstruction accuracy of real protein families, particularly for those widely relied-upon reconstructions [20–22]. The reconstruction of proteins from a single alignment of small numbers of orthologs may suffer from the same or worse accuracy issues we saw in single alignment approaches of our synthetic family. They may be worse in accuracy than what we observed in this study, since such reconstructions are derived from much smaller subsamples of ortholog sequences than we used in our subsample presented here and we found for small subsamples even for relatively high-Precision families individual reconstructions are extremely unreliable.

# Materials and Methods

## Preparation of synthetic sequence data

All sequence simulation materials and simulated sequence alignments are available via Figshare (10.6084/m9.figshare.5263885).

### Construction of phylogenies representing protein family divergence

Random 15-leaf phylogenies representing paralog divergence were generated at www.trex.uqam.ca [44] using the procedure of Kuhner and Felsenstein [45]. 100 phylogenies were generated with each average branch length of 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0, 600 in all. The Ensembl Compara species tree topology [46] containing 66 metazoan species was used for the divergence of each ortholog set. The topology was parametrized with branch lengths corresponding to species divergence times at http://www.timetree.org [47, 48]. For each of 15 leaves in each random phylogeny, a copy of the parametrized species tree was randomly scaled in overall height and had each individual branch randomly perturbed around its true length to maintain a realistic scale of divergence. The roots of these randomized trees (representing the MRCAs of an ortholog sets) were grafted at the leaves of the paralog phylogenies, resulting in 990-leaf synthetic protein family phylogenies.

### Preparation of sequence template and sequence simulation

Human tyrosine kinase domains were aligned using MAFFT L-INS-i with default parameters. This alignment was used as the template for sequence simulations as follows. The alignment was divided into 24 segments on the basis of local sequence similarity and analysis of solved tyrosine kinase structures. Each segment was assigned a substitution rate scaling factor and an insertion/deletion model to match degree of conservation and solvent exposure in solved structures. Simulation was carried out over synthetic phylogenies using indel-seq-gen [49–51] under the JTT substitution model.

## Phylogeny reconstruction

All-sequence phylogenies were inferred using all combinations of MAFFT L-INS-i, ClustalOmega, and Muscle for sequence alignment and of FastTree2 and RAxML for phylogeny inference. Subsampled phylogenies for Precision calculations (60 of 66 orthologs sampled from each ortholog set, 50 phylogenies reconstructed per protein family) were inferred with FastTree2 only, due to run time considerations. Subsampled phylogenies for ensembles (30 of 66 orthologs sampled, 1000 phylogenies per protein family) were reconstructed using L-INS-i and FastTree2 only.

Alignment algorithms were used with their default settings. FastTree2 was used with default settings and the WAG substitution model. RAxML was used with default settings and the PROTGAMMAWAGF variant of the WAG substitution model. The WAG substitution model was deliberately used for topology inference, instead of the JTT

substitution model used for simulating protein families, in order to emulate the more realistic scenario where models used for reconstruction of phylogenies for natural families do not precisely match the substitution patterns in those families.

Accuracy and Precision of reconstruction for a protein family are defined in terms of the L-INS-i / FastTree2 all-sequence and subsampled topology reconstructions.

### Modified Robinson-Foulds topology comparison metric

The Robinson-Foulds ($RF$) metric is defined in terms of leaf partitions at internal topology nodes for two topologies with identical sets of leaves. For a tree with $N$ leaves there are $N-3$ informative splits. The normalized form of the Robinson-Foulds comparison metric for two topologies, $A$ and $B$, is:

$$RF = \frac{x+y}{2N-6} \tag{1}$$

Where $x$ is the number of leaf partitions in A but not in B, y is the number of leaf partitions in B but not in A, N is the number of leaves in each topology, and $2N-6 = 2\times(N-3)$ is the number of informative splits in the two topologies.

In order to compare reconstructed paralog divergence topologies we had to modify the $RF$ metric to accommodate cases when the MRCA of an ortholog set has as descendants one or more MRCAs of other ortholog sets. Such topologies are poorly formed because they require inference of additional unobservable events – loss of paralogs in some lineages – in order to be reconciled with a duplication/speciation divergence history. Because the offending ortholog set cannot be pruned to a leaf MRCA, the resulting topology cannot be compared to properly formed topologies (e.g. the true topology) using the standard $RF$ metric. In effect, when ortholog leaves and speciation internal nodes of the offending ortholog set are pruned, the resulting topology is missing a MRCA leaf, because that MRCA maps to an internal node, making that node ambiguous in its duplication vs speciation status. This is problematic for $RF$ because it affects the denominator. Nevertheless, their internal nodes representing pre-duplication common ancestors of the offending ortholog set/paralog and other paralogs can match, in terms of induced partition of *paralogs*, equivalent nodes in other topologies.

In the modified $RF^*$, $N$ represents the number of paralogs (ortholog sets) in each compared topology, not the number of leaves. In addition to $x$ and $y$ we define $z$ as the number of MRCA leaves missing from $A$ but not from $B$ and $z'$ as the number of MRCA leaves missing from $B$ but not from $A$. The modified metric is then calculated as:

$$RF^* = \frac{x+y+z+z'}{2N-6} \tag{2}$$

### ASPEN

ASPEN is implemented in python 2.7. The ASPEN development repository is publicly available at *https://github.com/NaegleLab/ASPEN*.

## Acknowledgments

This work was enabled by the Center for High Performance Computing in the Mallinkrodt Institute of Radiology at Washington University in St. Louis and the Center for Biological Systems Engineering. We wish to thank Tom Ronan, Dr. Barak Cohen, Dr. Gary Stormo, Dr. Justin Fay, and Dr. Jim Havranek for the helpful discussions that shaped this work.

## References

1. Ohno, S. *Evolution by Gene Duplication* (Springer-Verlag, 1970).
2. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**, 309–38 (2005).
3. Prince, V. E. & Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3**, 827–37 (2002).

4. Raes, J. & Van de Peer, Y. Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. *Appl Bioinformatics* **2**, 91–101 (2003).

5. Espinosa-Cantú, A., Ascencio, D., Barona-Gómez, F. & DeLuna, A. Gene duplication and the evolution of moonlighting proteins. *Front Genet* **6**, 227 (2015).

6. Bielawski, J. P. & Yang, Z. *Maximum Likelihood Methods for Detecting Adaptive Protein Evolution*, chap. 5, 103–124. Statistics for Biology and Health (Springer New York, 2005).

7. Massingham, T. *Detecting the presence and location of selection in proteins.*, vol. 452 of *Methods in Molecular Biology*, chap. 15, 311–29 (United States, 2008).

8. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**, 908–17 (2002).

9. Massingham, T. & Goldman, N. Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753–62 (2005).

10. Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* **20**, 360–6 (2010).

11. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl 1**, S71–7 (2002).

12. Mihalek, I., Res, I. & Lichtarge, O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* **336**, 1265–82 (2004).

13. Dutheil, J. Y. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform* **13**, 228–43 (2012).

14. Thomson, J. M. *et al.* Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat Genet* **37**, 630–5 (2005).

15. Bridgham, J. T. *et al.* Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol* **8** (2010).

16. Eick, G. N., Colucci, J. K., Harms, M. J., Ortlund, E. A. & Thornton, J. W. Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet* **8**, e1003072 (2012).

17. Baker, C. R., Hanson-Smith, V. & Johnson, A. D. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* **342**, 104–8 (2013).

18. Rahman, T. *et al.* Two-pore channels provide insight into the evolution of voltage-gated ca2+ and na+ channels. *Sci Signal* **7**, ra109 (2014).

19. Creixell, P. *et al.* Unmasking determinants of specificity in the human kinome. *Cell* **163**, 187–201 (2015).

20. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–34 (2002).

21. Liu, B. A. *et al.* The human and mouse complement of sh2 domain proteins-establishing the boundaries of phosphotyrosine signaling. *Mol Cell* **22**, 851–68 (2006).

22. Chen, M. J., Dixon, J. E. & Manning, G. Genomics and evolution of protein phosphatases. *Sci Signal* **10** (2017).

23. Ogden, T. H. & Rosenberg, M. S. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology* **55**, 314–328 (2006).

24. Wong, K. M., Suchard, M. A. & Huelsenbeck, J. P. Alignment uncertainty and genomic analysis. *Science* **319**, 473–6 (2008).

25. Wang, L.-S. . S. *et al.* The impact of multiple protein sequence alignment on phylogenetic estimation. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4, 1108–1119. University of Texas at Austin, Austin (IEEE Computer Society, 2009).

26. Liu, K., Linder, C. R. & Warnow, T. Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Curr* **2**, RRN1198 (2010).

27. Blackburne, B. P. & Whelan, S. Class of multiple sequence alignment algorithm affects genomic analysis. *Mol Biol Evol* **30**, 642–53 (2013).

28. Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. & Warnow, T. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324**, 1561–4 (2009).

29. Landan, G. & Graur, D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol* **24**, 1380–3 (2007).

30. Wu, M., Chatterji, S. & Eisen, J. A. Accounting for alignment uncertainty in phylogenomics. *PLoS One* **7**, e30288 (2012).

31. Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**, 0126 (2017).

32. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–31 (2013).

33. Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

34. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol* **7**, 539 (2011).

35. Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–7 (2004).

36. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

37. Stamatakis, A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–3 (2014).

38. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical biosciences* **53**, 131–147 (1981).

39. Morrison, D. A. & Ellis, J. T. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18s rdnas of apicomplexa. *Mol Biol Evol* **14**, 428–41 (1997).

40. Mugridge, N. B. *et al.* Effects of sequence alignment and structural domains of ribosomal dna on phylogeny reconstruction for the protozoan family sarcocystidae. *Molecular Biology and Evolution* **17**, 1842–1853 (2000).

41. Liu, K., Linder, C. R. & Warnow, T. Raxml and fasttree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One* **6**, e27731 (2011).

42. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res* **45**, D635–D642 (2017).

43. Benson, D. A. *et al.* Genbank. *Nucleic Acids Res* **45**, D37–D42 (2017).

44. Boc, A., Diallo, A. B. & Makarenkov, V. T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res* **40**, W573–9 (2012).

45. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* **11**, 459–68 (1994).

46. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database (Oxford)* **2016** (2016).

47. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* **32**, 835–45 (2015).

48. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. Timetree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34**, 1812–1819 (2017).

49. Strope, C. L., Scott, S. D. & Moriyama, E. N. indel-seq-gen: a new protein family simulator incorporating domains, motifs, and indels. *Mol Biol Evol* **24**, 640–9 (2007).

50. Strope, C. L., Abel, K., Scott, S. D. & Moriyama, E. N. Biological sequence simulation for testing complex evolutionary hypotheses: indel-seq-gen version 2.0. *Mol Biol Evol* **26**, 2581–93 (2009).

51. Rambaut, A. & Grassly, N. C. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comput Appl Biosci* **13**, 235–8 (1997).