CHAPTER

Reconciliation Approaches to Determining HGT, Duplications, and Losses in Gene Trees

9

Olga K. Kamneva[*,1], **Naomi L. Ward**[†]

[*]*Department of Biology, Stanford University, Stanford, California, USA*

[†]*Department of Molecular Biology, University of Wyoming, Laramie, Wyoming, USA*

[1]*Corresponding author: e-mail address: okamneva@stanford.edu*

1 INTRODUCTION

Dramatic gene loss and gain, corresponding to physiology and lifestyle shifts, has been documented for a number of bacterial lineages (Lefébure & Stanhope, 2007). Such changes are of interest to a variety of disciplines, including microbiology, molecular and cellular biology, medicine, and biotechnology. Changes in genomic content occur due to evolutionary processes such as gene duplication, loss, and horizontal gene transfer (HGT). These events lead to differences between the histories of different genes, as well as discrepancies between the evolutionary history of individual genes and the genomes overall (Pamilo & Nei, 1988). Analyses of different gene family histories, or *genealogies*, can also be used to infer the kinds of events that have taken place in the evolutionary history of a genome, providing insight into the ways in which genomes have changed over time, including functional changes (Kamneva, Knight, Liberles, & Ward, 2012).

Explaining inconsistencies between the evolutionary histories of genes and the species in which they evolve is called *reconciliation*. Reconciliation has a number of important practical applications. For instance, reconciliation is the most comprehensive way to describe the dynamics of gene family evolution in terms of gene copy number (David & Alm, 2011; Kamneva et al., 2012). It is also the most reliable way to identify truly orthologous genes between different genomes (Åkerborg, Sennblad, Arvestad, & Lagergren, 2009), which is important when using information about a gene in one organism to understand the function of related genes in other organisms, an important part of genome annotation and sequence analysis.

Therefore, comparisons between the family histories of genes and the history of the species in which they evolve are becoming a common practice in microbiology research. The aims of these comparisons are to (1) predict the functions and

Methods in Microbiology, Volume 41, ISSN 0580-9517, http://dx.doi.org/10.1016/bs.mim.2014.08.004

properties of newly characterized genes and genomes, (2) characterize the evolutionary history of individual genes, (3) characterize genome evolution in terms of gene family content, and (4) predict ancestral gene family composition.

In this chapter, we describe how to apply some of the methods developed for gene and genome history reconciliation. We include several real data analysis examples to illustrate the different reconciliation techniques, and we include explicit analysis protocols with every example.

2 BACTERIAL SPECIES TREE

A *biological species* is a group of genetically similar organisms that are capable of interbreeding and producing fertile offspring. The evolutionary relationships among different species are represented by a tree-like graph called a "species tree," in which every branching point represents the divergence of one species into two or more new species. It is generally assumed that each divergence, or *speciation event*, occurs at a fixed point in time (Figure 1A). However, in bacteria, the classical definition of a species is challenged because gene flow can occur between closely related strains via recombination during the time of species separation, and between distantly related species via HGT (for a review, see Ochman, Lawrence, & Groisman, 2000; Retchless & Lawrence, 2010). Processes of hybridization and HGT lead to different evolutionary histories in different parts of a genome. Therefore, in any particular set of species, the classical model of species evolution (a bifurcating tree) is somewhat invalidated (McInerney, Pisani, Bapteste, & O'Connell, 2011). The accumulating evidence for non-tree-like patterns in the evolutionary histories of different organisms has led to the view that species histories are best represented by networks, rather than trees (Figure 1B). However, the implementation of software tools based on this new view of species evolution is still under development. Many approaches for genomic analysis relying on the assumption that horizontal processes (recombination, HGT, hybridization) occur in the context of a species history that is fundamentally tree-like often lead to satisfactory results in molecular evolutionary studies (David & Alm, 2011; Kamneva et al., 2012).

The evolutionary history of a set of species, or the *species tree*, is one of the necessary components of the analyses described in this chapter. Several methods are currently used to infer species trees from DNA and protein sequences. Classical approaches in molecular taxonomy utilize sequences of universally distributed genes

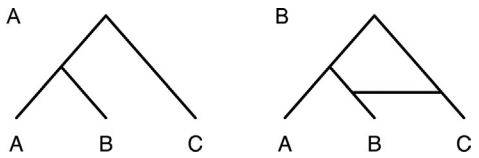


FIGURE 1

Example species tree (A) and network (B).

encoding the rRNA of the small subunit of the ribosome (16S rRNA in bacteria and archaea and 18S rRNA in eukaryotes). Since the 1970s, when 16S rRNA genes were first used in molecular phylogenetics (Woese & Fox, 1977), comprehensive protocols for 16S rRNA sequencing and further data processing have been developed (see Ludwig, Oliver Glöckner, & Yilmaz, 2011 for further details). However, the limited phylogenetic information available within one molecular marker often leads to unresolved or poorly supported phylogenies. This difficulty has been addressed by using multiple universally distributed protein-coding genes for phylogeny reconstruction (Rokas, Williams, King, & Carroll, 2003). The development of whole-genome sequencing and re-sequencing projects has provided data that are well suited for multi-locus species tree reconstruction.

Several software tools for gene and genome history reconciliation also require dated species trees, in which every node in the tree has a date assigned to it and the branch lengths of the tree are measured in units of time, rather than in the numbers of nucleotide or amino acid substitutions occurring along each branch. The protocol to produce such a tree is described and utilized in David and Alm (2011), Kamneva et al. (2012), and Parfrey, Lahr, Knoll, and Katz (2011) and relies on the assumption of a molecular clock, whereby the amount of mutational change is a linear function of time, which is calibrated using dated archaeological fossils. This type of analysis is complicated and time consuming; therefore, the use of an existing dated species trees from previous analyses is often accepted (Ciccarelli et al., 2006).

3 GENE FAMILY

Another central component of gene–genome history reconciliation is the gene family, a group of sequences in different genomes that share a common history and which presumably carry out the same functions. Homologous gene families are constructed in different ways, described in detail in Pertea et al. (2003) and Tatusov et al. (2003). The process mostly relies on sequence comparison between genes from the same or different genomes, following the clustering of genes into different gene families (Li, Stoeckert, & Roos, 2003).

Two canonical representations of gene family are the gene tree (Figure 2A) and the phyletic pattern or profile (Figure 2B). A phyletic profile is a very simple

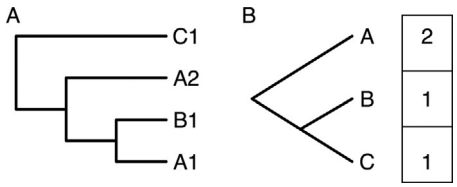


FIGURE 2

Phylogenetic tree representing evolution of a gene family (A) and its phylogenetic profile representation (B).

representation of gene family. It conveys information only about the presence or absence of a gene in contemporary genomes (Figure 2B). On the other hand, a phylogenetic gene tree reflects how sequences of different genes sampled from extant organisms are related to each other through various evolutionary events, such as gene duplication, HGT, speciation, and divergence (Figure 2A). The phylogenetic gene tree provides more information about the gene family. The reconstruction of the gene tree phylogeny is a challenging task on its own, especially when very closely or very distantly related sequences are considered. An example protocol for reconstructing gene phylogenies using either protein or DNA sequences is described here (Rokas, 2001) and include the following logical steps: (1) collecting sequences belonging to the gene family of interest; (2) aligning sequences in the data set to each other to generate a multiple sequence alignment; (3) testing for the evolutionary model of best fit; and (4) building a gene phylogeny.

It is important to identify gene families accurately, as including distantly related paralogous genes as members of the same family will lead to a number of problems. In the case where a phylogenetic profile is used to represent a gene family, additional duplication events might be predicted and erroneously placed onto lineages of the species tree. Including distantly related genes will also cause problems with gene phylogeny inference due to, for instance, long-branch attraction artefacts (Bergsten, 2005). Over-prediction of duplication and loss events or HGT events, and erroneous placement of the events over species tree lineages are also possible in this case. On the other hand, excluding some sequences from the gene family is also problematic, as then the gene family might be predicted to originate on the wrong lineage of the species tree, or additional gene loss or HGT events might be erroneously inferred.

4 EVOLUTION OF GENES IN BACTERIAL GENOMES

Multiple evolutionary processes affect genes in bacterial genomes over time. The set of events and processes that affect a gene family constitute its evolutionary history. Ultimately, the history of each gene should be considered in light of the history of the genome. As we discussed previously, speciation is a significant event in genome evolution (Figure 3A). Genes that originate from a common ancestor and that have been separated by speciation events are called orthologous genes (Figure 3B).

Other evolutionary processes include gene duplication, recombination, HGT, and gene loss. Gene duplication occurs when a second copy of a gene is created in a genome. Genes separated by an ancestral duplication event are called paralogs (Figure 3C). In the simplest case, duplication can be inferred from the gene history when two clades in the gene tree closely resemble the complete species tree or a part of it. This resemblance is the result of the fact that, after a duplication event, both copies of the gene follow similar evolutionary paths along the species tree. Different orders in the events of duplication and speciation give rise to in- and out-paralogs (Figure 3C). Divergence after gene duplication might lead to the emergence of more

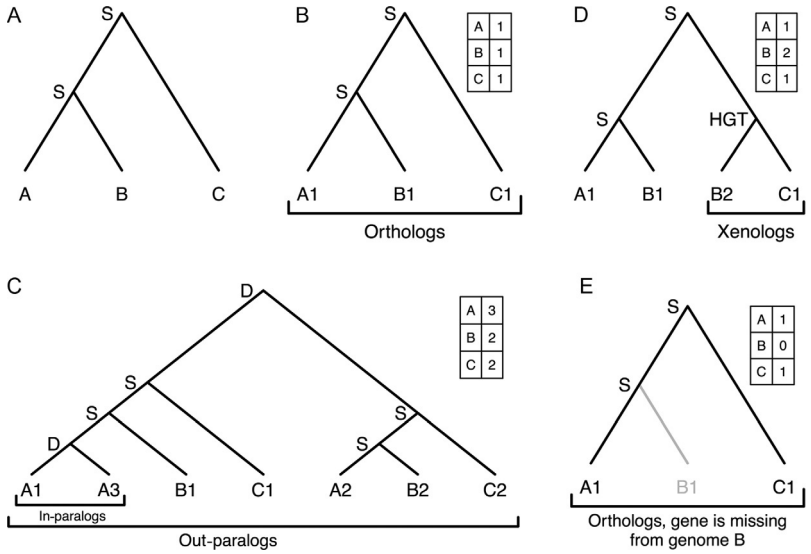


FIGURE 3

Species tree (A), gene tree relating orthologs and corresponding phyletic pattern (B); gene tree relating in-, out-paralogs and orthologs and corresponding phyletic pattern (C); gene tree relating xenologs and orthologs and corresponding phyletic pattern (D); gene tree relating orthologous genes and corresponding phyletic pattern; gene loss has removed the gene from genome B (E). Events are denoted as follows: S, speciation; D, duplication; HGT, horizontal gene transfer.

specialized or slightly different biological functions within the paralogous gene pairs. A set of orthologs and in-paralogs represents a gene family in a classic orthology analysis used in comparative genomics-based gene function prediction. The converse of gene duplication is gene loss. Gene loss can be detected in the gene history due to missing lineages observed in a gene tree but present in the species tree (Figure 3E).

When characterizing gene evolution in bacteria, it is important to account for recombination and HGT as well as other events. Recombination is a process that affects an ensemble of recombining bacterial strains that are often associated with distinct ecological niches, but which nevertheless maintain a sufficient amount of genetic similarity to recombine with one another.

With time, bacterial lineages diverge and genetic isolation will be established for the major parts of the genome, although this does not completely prohibit the further exchange of genetic information. HGT is often observed between distantly related bacterial genomes. It can be facilitated by mobile genetic elements and bacteriophages after transformation, conjugation, or transduction. Different kinds of mobile elements are also major forces driving genomic rearrangements and allowing for high levels of genomic plasticity. Genes arising from HGT events in the gene history

are called xenologs (Figure 3D). Xenologs can be detected in the gene phylogeny, as they will cluster with the genes to which they are more closely related from organisms that are often distantly related.

In reality, the number of events occurring on different genetic lineages at different times might confound patterns of duplication, loss, and HGT. Thus, rigorous methods are needed in order to characterize gene family evolution in light of these large-scale evolutionary processes. A variety of such methods currently exist. They reconcile detailed gene trees or gene family-specific presence/absence profiles with a species tree through the inference of gene gain, loss, duplication, and HGT events. Methods that can perform inference of HGT events are particularly relevant for the analysis of bacterial genome evolution.

It is important to point out that alternative scenarios of gene family evolution exist. Ideally, one should evaluate which scenario is the most likely one; however, the majority of currently used, parsimony-based, methods do not allow such evaluation and require the user to rely on very arbitrary measures. Therefore, it is sometimes challenging to evaluate how realistic any given inferred gene history is.

5 GENE TREE/SPECIES TREE RECONCILIATION

In essence, gene tree/species tree reconciliation procedures aim to explain inconsistencies between gene and genome history using various evolutionary events. Figure 4 illustrates this process. It is important to note the existence of a number of alternative scenarios consistent with the given species and gene history. While running the analysis it is important to be able to choose the most plausible scenario. For that a number of software tools to analyze gene evolution in the context of species trees have been developed (Table 1). Various factors should be considered when performing such analyses. One important component of such an analysis is an accurate estimate of the species tree. The majority of existing methods for gene tree/species tree reconciliation require fully resolved, bifurcating, species trees which are sometimes hard to generate. However, several well-resolved and sufficiently supported species trees for a wide range of organisms have been reported and are sometimes used (Ciccarelli et al., 2006).

Another important thing to consider is the statistical framework that is used. Generally, Bayesian and likelihood-based methods allow reliable estimation of model parameters; however, they generally take a long time to run, which can be problematic in the case of large data sets or genome-level studies. In addition, many existing tools that use Bayesian or likelihood approaches do not consider lateral transfer events, which can be an especially important point to consider in the case of bacteria. Some parsimony-based methods do address transfer events, which makes them very useful for bacterial comparative genomics. They use some kind of event penalties to evaluate prospective evolutionary scenarios. The more events occur, the less parsimonious and less favourable the scenario is taken to be.

Representation of the gene family in use is another factor to take into account. A gene tree provides a more detailed description of a gene family than a phyletic

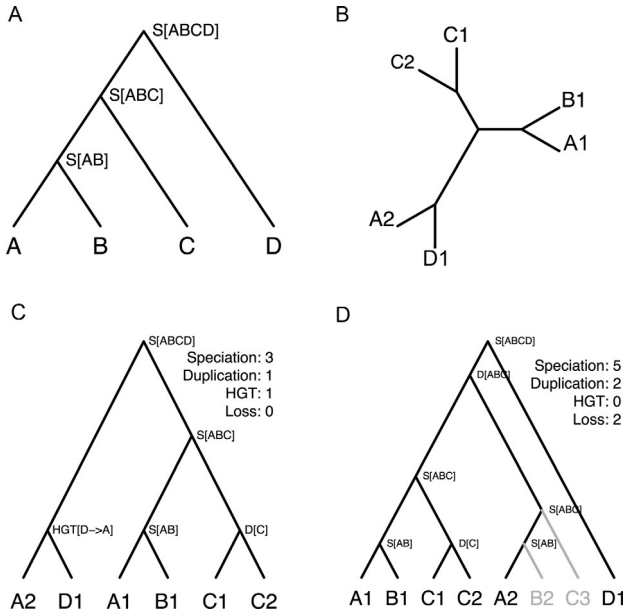


FIGURE 4

Species tree, with speciation events denoted by the letter S and the name of the emerging lineage given in squared brackets (A); unrooted gene tree (B); one plausible scenario of gene family evolution consistent with the species relationships (C); alternative scenario (D). Speciation, duplication, and HGT events are marked by letters S, D, and HGT sign, ancestral lineages where events are predicted to occur are indicated in squared brackets. Gene loss is indicated by a grey colour of the lineage and corresponding leaves.

profile. The use of phyletic patterns can be problematic, especially when analyzing genes from distantly related organisms because duplications, losses, and transfers that occur over long time periods could cancel each other out and they might not be reflected in the phyletic profile. On the other hand, gene tree reconciliation methods that use detailed gene histories should be robust to errors in gene tree inference. This is especially relevant for analyses involving distantly, or very closely, related sequences.

Another issue to consider is the taxonomy of the organisms included in the analysis. It is appealing to consider large number of genomes to obtain the most comprehensive view of gene evolution. However, including many species in analyses often results in long computation time. This is especially relevant in cases of whole-genome analyses. It is also important to make sure that taxonomic sampling is not biased towards one or several taxonomic groups or species with certain lifestyles. It is considered the best practice to include evolutionarily and taxonomically divergent species associated with various ecological habitats in the analysis.

Table 1 Some Programs for Gene Tree Species/Tree Reconciliation and Species Tree-Guided Gene Tree Reconstruction

Program	Events Considered	Framework	Time Consistent Transfers	Input
AnGST (David & Alm, 2011)	Duplication Loss HGT	Parsimony	Yes (if dated species tree is provided)/ No	Bifurcating species tree One or several gene trees Event penalties
MPR (Doyon, Hamel, & Chauve, 2010)	Duplication Loss HGT	Parsimony	Yes	Dated, bifurcating species tree One or several gene trees Event penalties
Ranger-dtl (Bansal, Alm, & Kellis, 2012)	Duplication Loss HGT	Parsimony	No	Bifurcating species tree One or several gene trees Event penalties
DLCpa (Wu, Rasmussen, Bansal, & Kellis, 2014)	Duplication Loss ILS	Parsimony	No	
Prime dltrs (Åkerborg et al., 2009)	Duplication Loss HGT	Bayesian	Yes	Dated, bifurcating species tree Gene sequence alignment Model of evolution

Here, we describe protocols for running AnGST, a parsimony-based program for gene tree–species tree reconciliation which requires python as well as basic familiarity with a command-line interface. AnGST was introduced by David and Alm (2011), and it can be obtained from http://almlab.mit.edu/angst/files/angst.tar.gz or run *via* the web application at http://almlab.mit.edu/angst/. AnGST implements an algorithm that uses gene birth, duplication, losses, and HGT to explain discrepancies between gene trees and species trees, assuming that evolutionary events that lead to discordance between a gene tree and the species tree carry fixed costs. Here, we examine the evolution of DUF70 domain proteins from several archaeal genomes

using a locally installed version of AnGST. The protocol for running the same analysis on the server does not differ significantly.

The costs of evolutionary events can potentially be set to arbitrary values, which have significantly different effects on the inference. Here, we will examine the effect that penalty values have on the results by conducting the inference using various event costs.

AnGST also allows the user to incorporate information about gene tree uncertainty into the analysis. This is done via a bootstrap amalgamation procedure when the gene tree with the lowest reconciliation cost is chosen from a collection of trees that are consistent with the set of bipartitions present in all of the input gene trees.

5.1 PROTOCOL FOR RUNNING AnGST

1. Download the online supplementary files from http://dx.doi.org/10.1016/bs.mim.2014.08.004. File AnGST_ST contains a dated, bifurcating species tree for 100 organisms. It can be viewed using a number of tree viewers, for instance FigTree. File AnGST_GTs contains a set of gene trees from several bootstrap runs (this file might also contain a single gene tree). Individual gene trees can also be viewed using any phylogeny visualization tool. AnGST_penalties is a penalties file where penalties for birth, duplication, loss, and HGT events are defined. AnGST_run1.input and AnGST_run2.input are control files for AnGST. Familiarize yourself with these files. Note that gene naming follows a certain convention where gene names include species names separated by the symbol ".". However, other symbols can be used as well (see the AnGST manual for details).
2. Create a directory (folder) where you want your results to go, and place all of the files (AnGST_ST, AnGST_GTs, AnGST_costs) within it. Now open a terminal and move to the directory that contains those files. Assuming that AnGST is in the directory located at /path_to_AnGST/, run the program using following command in the command-line prompt:
   ```
   /path_to_AnGST/python angst_lib/AnGST.py AnGST_rin1.input
   ```
3. Familiarize yourself with the results (they are also available at AnGST_run1.zip). If you have not edited the control file, the results will be written to a directory called AnGST_run1. The AnGST.nexus file contains a single resulting gene tree with every node annotated with the genome name in which it is predicted to be present. Additionally, the AnGST.events file contains a list of all the events asserted to have happened in the history of the gene under consideration.
4. Now use a text editor to change the event costs within the AnGST_costs file and re-run the program. The objective here is to compute the reconciliation for the example gene family, given different sets of event costs, and observe how predicted evolutionary scenario changes with the change of event penalties.
   a. Use HGT cost equal to 1, leaving all the other penalties unchanged
   b. Run AnGST again using AnGST_run2.input control file:
      ```
      /path_to_AnGST/python angst_lib/AnGST.py AnGST_rin2.input
      ```
5. Examine results of two additional runs, pay special attention to AnGST.events file, and note varying number of different events inferred for the data set.

5.2 INTERPRETING THE RESULTS OF AnGST ANALYSES

Results of these three AnGST runs are also visualized in Figure 5. It is clear that they differ quite a bit. When the penalty for HGT decreases relatively to the cost of duplication, more transfer events are inferred. This on its own indicates that every possible reconciliation can potentially be recovered with varying event penalties under parsimonious inference. This calls for additional justification of event cost values. In the original article introducing AnGST, the authors used genome flux (average change in genome size over every lineage of species tree) analysis to justify duplication, loss, and transfer costs (David & Alm, 2011). They obtained values included in the original AnGST_costs file. Very similar values for event penalties were recovered using genome flux analysis in a different study, on the data set including only bacterial genomes (Kamneva et al., 2012). Therefore, since optimization of event penalties using genome flux is computationally intensive, use of previously reported event penalties might be acceptable.

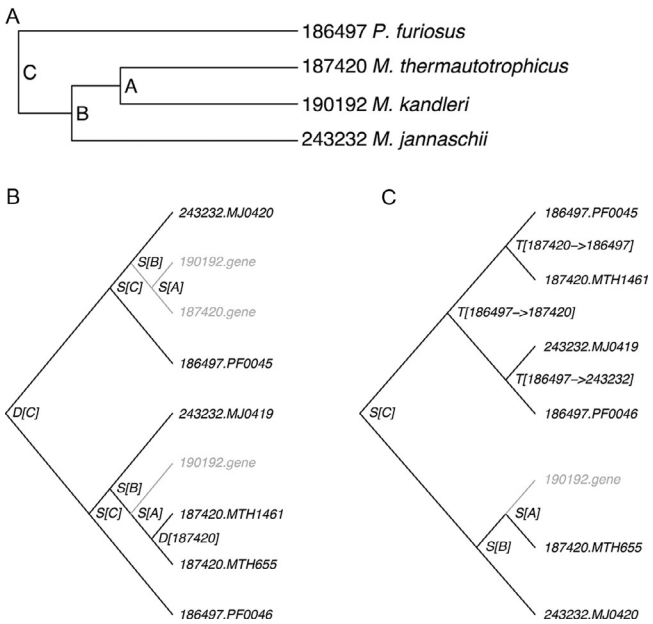


FIGURE 5

Species tree, names of ancestral genomes are shown as node labels and NCBI taxonomy IDs are given as tip labels along with the abbreviated species name (A); one plausible scenario of gene family evolution consistent with the species relationships as inferred in the first run of AnGST (B); alternative scenario, as inferred in the second AnGST run (C). Speciation, duplication, and HGT events are marked by letters S, D, and T, ancestral lineages where events are predicted to occur are indicated in squared brackets. Gene loss is indicated by a grey colour of the lineage and corresponding tip labels.

6 ANALYSIS AT THE GENOME SCALE

Whole-genome-level analysis has become very popular as a result of the accumulation of a large number of sequenced bacterial genomes. Here, we illustrate how this type of analysis is performed using the COUNT program, as it was designed to work on genome-scale data sets. However, genome-level analyses can be performed on a family-by-family basis using one of the protocols described in the previous sections.

The COUNT software tool was introduced by Csurös and Miklós (2009) and later by Csűös (2010). It is available from the authors' website: http://www.iro.umontreal.ca/∼csuros/gene_content/count.html. It implements a variety of methods for evolutionary analysis of phyletic patterns or other types of integer-valued evolutionary characters, including parsimony, and probabilistic methods based on a phylogenetic birth-and-death model. The latter functionality is the most interesting within the framework of this chapter.

Ignoring explicit gene phylogenies within COUNT allows one to fit more sophisticated models of gene family evolution and estimate model parameters using genome-wide data. The implemented birth-death model assumes gene loss, duplication, and gain occurs along each branch of the species tree according to a stochastic process. Gene losses, gains, and duplications are assigned gene family- and edge-specific rates (k_f, k_e, lambda$_f$, lambda$_e$, mu$_f$, mu$_e$), and the processes run along the species tree edges for the time $t_e * t_f$. Edge-specific parameters are either held constant or vary along the branches of the species tree. Gene family-specific parameters are also assumed to be either constant or have a discretized Gamma distribution.

6.1 PROTOCOL FOR RUNNING COUNT

Here, we provide a protocol for running probabilistic inference of genome content using COUNT:

1. Obtain the online supplementary files from http://dx.doi.org/10.1016/bs.mim.2014.08.004. The file COUNT_ST contains a species tree for five archaeal organisms. It can be viewed using any phylogeny visualization program. The page COUNT_table contains a list of gene families in the form of phyletic patterns and can be viewed in any text editor or Microsoft Excel. Familiarize yourself with these files. The file COUNT_commands contains command-prompt lines to be used to execute COUNT. The file content can be viewed using text editor.
2. Create a directory where you want your results to go, and place both data files (COUNT_ST, COUNT_table) within it. Now open a terminal and move to the directory that contains those files. Assuming that count.jar is in the directory located at /PATH_TO_COUNT/, run the analysis using commands included in the COUNT_commands file. This will allow you to fit a general model of genome evolution in a hierarchical fashion from the simplest to more and more complex.

3. Examine the resulting rate files, which are also enclosed within the COUNT_results.zip. Identify the line within the results file that gives the likelihood score for the data set. It is denoted as final likelihood but it appears as the negative logarithm of the likelihood score.
4. In the next step, we would like to compare models to each other to find the least complex model that fits the data reasonably well. We will use AIC for this purpose. A general review of this topic is provided by Burnham and Anderson (2002). In order to employ AIC, knowledge of the number of parameters estimated by the model in every COUNT run is required (here we denote the model with the name in the generated rate file):
   run1.rates: 3 edge-specific parameters + 1 parameter for gene family size distribution at the root of the species tree = 4 = k
   run2.rates: 3 * 8 edge-specific parameters + 1 parameter for gene family size distribution at the root of the species tree = 25 = k
   run3.rates: 3 * 8 edge-specific parameters + 1 parameter for gene family size distribution at the root of the species tree + 1 parameter for Gamma distribution for family-specific branch length adjustment factor = 26 = k
   run4.rates: 3 * 8 edge-specific parameters + 1 parameter for gene family size distribution at the root of the species tree + 1 parameter for Gamma distribution for family-specific branch length adjustment factor + 1 parameter for Gamma distribution for family-specific duplication rate = 27 = k
   run5.rates: 3 * 8 edge-specific parameters + 1 parameter for gene family size distribution at the root of the species tree + 1 parameter for Gamma distribution for family-specific branch length adjustment factor + 1 parameter for Gamma distribution for family-specific duplication rate + 1 parameter for Gamma distribution for family-specific loss rate = 28 = k
   run6.rates: 3 * 8 edge-specific parameters + 1 parameter for gene family size distribution at the root of the species tree + 1 parameter for Gamma distribution for family-specific branch length adjustment factor + 1 parameter for Gamma distribution for family-specific duplication rate + 1 parameter for Gamma distribution for family-specific loss rate + 1 parameter for Gamma distribution for family-specific gain rate = 29 = k
   Use Microsoft Excel or other tool to calculate AIC as $2k - 2\ln(L)$. Model number 6 (run6.rates) with the lowest AIC value (42,283.77) should be recognized as the best fitting one.
5. Now in order to infer the ancestral composition of gene families for every node of the species tree and characterize the evolutionary dynamics of every gene family, we will use the COUNT application which has a graphical interface and allows both posterior analysis and data visualization. Start the COUNT application obtained with the COUNT package. Choose *new session* option under *Session* menu, load the species tree from COUNT_ST file; choose *open table* option under *Data* menu, load the gene family table from COUNT_table file; and lastly, choose *load rates* option under *Rates* menu from run6.rates file. To run the analysis, select *family history by posterior probability* option within the *Analysis*

menu. Alternatively, to obtain data for further manipulation and visualization using custom tools, the Posterior application can be run with the following syntax:

```
java -Xmx2048M -cp /PATH_TO_COUNT/Count.jar
ca.umontreal.iro.evolution.genecontent.Posteriors -max_paralogs m
OUNT_ST COUNT_table run6.rates > run6.posteriors
```

The run6.posteriors file will contain information on the probability of gene family (zero, one, or multiple genes are predicted to be present at every node of gene tree) and dynamics (gene family expansion, contraction, loss, or gain is predicted) at every node of the species tree for every gene family in the data set. For simplicity, here we proceed with the GUI version of the workflow.

After the analysis is finished, gene family dynamics can be viewed under the *Data* tab within the application. Family names will be shown along with the information on family size, number of extant lineages containing at least one member of the gene, and phyletic pattern, and information on gene family composition and dynamics at every node of the species tree will also be shown for every gene family as well (Figure 6A). Information about every gene family can be viewed individually for every gene family in tree display (Figure 6B) or in a tabular format (Figure 6C). The ArCOG00175 family is predicted to have been present in the ancestor of all the class *Methanomicrobia* as a multigene family and then to have been contracted and lost on the lineage leading to *Methanococcoides burtonii* DSM 6242.

Additionally, a summary across the species tree for a set of gene families can be obtained by selecting desired gene families from the data table. Figure 7 shows the genome-wide summary obtained by selecting all the gene families, visualized over the tree (Figure 7A) and in the tabular format (Figure 7B). It appears that on the lineage leading to the ancestor of all species in the order Methanomicrobiales, some gene families were gained while a lot more lost, and twice as many gene families were predicted to shrink in size compared to expanded families. Both trends are consistent with genome contraction happening on the lineage.

CONCLUDING REMARKS

The field of comparative and evolutionary genomics is developing fast, driven by the availability of genome sequence data. However, new techniques and analysis protocols should be used with caution. With complicated multi-stage analyses, errors and biases can be introduced at any step, leading to systematic biases in the results and conclusions. Additionally, benchmarking studies are rarely published alongside the methods for gene tree/species tree reconciliation, which constitutes a significant knowledge gap and calls for those benchmarking studies.

In this chapter, we have discussed issues that should be considered when performing gene tree/species tree reconciliation to identify duplications, losses, and transfers. They include (1) the nature of the taxonomic and ecological sampling of

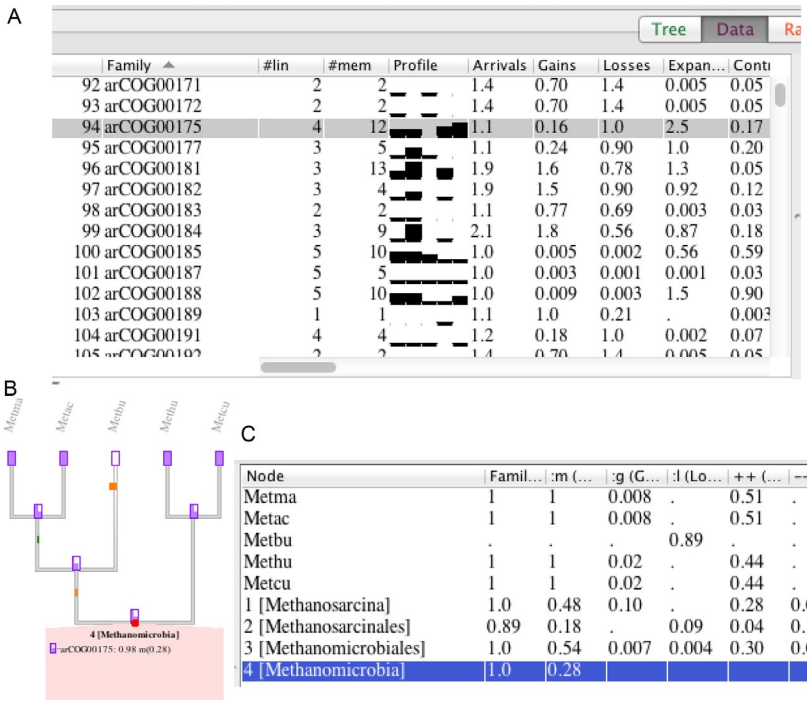


FIGURE 6

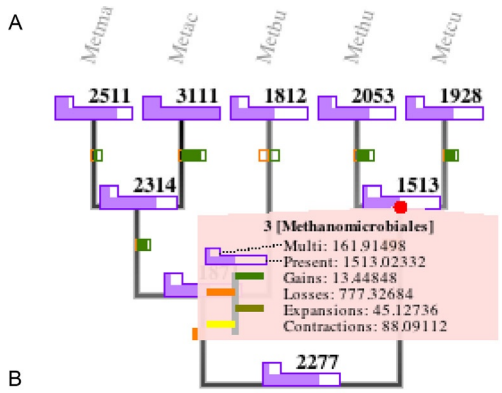
Gene family evolution as inferred by COUNT, general information across the families (A); and gene family history for arCOG00175 over the species tree (B) and in tabular format (C). Species are indicated as follows: Methma (*Methanosarcina mazei*), Methac (*Methanosarcina acetivorans*), Methbu (*Methanococcoides burtonii* DSM 6242), Methu (*Methanospirillum hungatei* JF-1), and methma (*Methanoculleus marisnigri* JR1).

genomes included in the analysis; for instance, including taxonomically divergent but ecologically similar species in the data set might lead to underestimation of the number of HGT events within a gene family of niche-specific genes; (2) errors in gene family identification; for instance, inclusion of distant paralogs not only complicates reconciliation but complicates inference of gene tree topology; (3) uncertainties in gene and species tree estimation; and (4) choice of computer program or algorithms with realistic underlying assumptions and justified parameter values. The last point is, in our opinion, the most influential one.

Some additional points that have not been discussed here include the inference of duplication, loss, and transfer events using gene order/synteny. This practice seems to lead to good results on empirical data, but protocols for such studies are not established.

One additional phenomenon that is widely acknowledged in other fields of evolutionary biology, but rarely discussed by empiricists in the gene tree/species tree

A



B

Node	Famil...	:m (...	:g (G...	:l (Lo...	++ (...	-- (C...
Metma	2511	551	316	119	168	25
Metac	3111	737	948	151	278	28
Metbu	1812	318	357	416	124	55
Methu	2053	520	683	143	222	14
Metcu	1928	389	522	107	170	13
1 [Methanosarcina]	2314	384	493	50	130	11
2 [Methanosarcinales]	1871	228	22	429	68	77
3 [Methanomicrobiales]	1513	162	13	777	45	88
4 [Methanomicrobia]	2277	258				

FIGURE 7

Genome content evolution as inferred by COUNT, genome-wide summary is shown over the species tree and in tabular format. Species are indicated as follows: Methma (*Methanosarcina mazei*), Methac (*Methanosarcina acetivorans*), Methbu (*Methanococcoides burtonii* DSM 6242), Methu (*Methanospirillum hungatei* JF-1), and methma (*Methanoculleus marisnigri* JR1).

reconciliation community, is incomplete lineage sorting (ILS). ILS is the discordance between gene trees and species trees due to stochasticity in the coalescence times of ancestral genetic lineages. In a number of studies, a large amount of HGT has been identified between closely related organisms; however, much of the signal for HGT could in fact be due to ILS, due to the large effective population sizes in bacteria. Although some methods exist for inferring horizontal transmission of genetic information while accounting for ILS (Yu, Barnett, & Nakhleh, 2013), they do not treat duplication and loss events and the methodology in this area is still very much under development.

ACKNOWLEDGEMENTS

The authors thank Ethan Jewett and Miklós Csurös for helpful discussions. This work was supported by The Stanford Center for Computational, Evolutionary and Human Genomics

Postdoctoral Fellowship and National Science Foundation (MCB-0920667 to N. L. W. and DBI-1146722 to O. K. K.). The content of this chapter is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

REFERENCES

Åkerborg, Ö., Sennblad, B., Arvestad, L., & Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5714–5719. http://dx.doi.org/10.1073/pnas.0806251106.

Bansal, M. S., Alm, E. J., & Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, *28*(12), i283–i291. http://dx.doi.org/10.1093/bioinformatics/bts225.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163–193. http://dx.doi.org/10.1111/j.1096-0031.2005.00059.x.

Burnham, K., & Anderson, D. R. (2002). *Model selection and multimodel inference—A practical information-theoretic approach*. Retrieved from, http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95364-9.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, *311*(5765), 1283–1287. http://dx.doi.org/10.1126/science.1123061.

Csűös, M. (2010). Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, *26*(15), 1910–1912. http://dx.doi.org/10.1093/bioinformatics/btq315.

Csurös, M., & Miklós, I. (2009). Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Molecular Biology and Evolution*, *26*(9), 2087–2095. http://dx.doi.org/10.1093/molbev/msp123.

David, L. A., & Alm, E. J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, *469*(7328), 93–96. http://dx.doi.org/10.1038/nature09649.

Doyon, J.-P., Hamel, S., & Chauve, C. (2010). *An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework*. Retrieved from, http://hal-lirmm.ccsd.cnrs.fr/lirmm-00448486.

Kamneva, O. K., Knight, S. J., Liberles, D. A., & Ward, N. L. (2012). Analysis of genome content evolution in PVC bacterial super-phylum: Assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biology and Evolution*, *4*(12), 1375–1390. http://dx.doi.org/10.1093/gbe/evs113.

Lefébure, T., & Stanhope, M. J. (2007). Evolution of the core and pan-genome of *Streptococcus*: Positive selection, recombination, and genome composition. *Genome Biology*, *8*(5), R71. http://dx.doi.org/10.1186/gb-2007-8-5-r71.

Li, L., Stoeckert, C. J., Jr., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. http://dx.doi.org/10.1101/gr.1224503.

Ludwig, W., Oliver Glöckner, F., & Yilmaz, P. (2011). *16—The use of rRNA gene sequence data in the classification and identification of prokaryotes*. In F. Rainey & A. Oren (Eds.), *Methods in microbiology: Vol. 38* (pp. 349–384). Waltham, MA, USA: Academic Press. Retrieved from, http://www.sciencedirect.com/science/article/pii/B9780123877307000164.

McInerney, J. O., Pisani, D., Bapteste, E., & O'Connell, M. J. (2011). The Public Goods Hypothesis for the evolution of life on earth. *Biology Direct*, *6*, 41. http://dx.doi.org/10.1186/1745-6150-6-41.

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*(6784), 299–304. http://dx.doi.org/10.1038/35012500.

Pamilo, P., & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, *5*(5), 568–583.

Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., & Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13624–13629. http://dx.doi.org/10.1073/pnas.1110633108.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., et al. (2003). TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics*, *19*(5), 651–652. http://dx.doi.org/10.1093/bioinformatics/btg034.

Retchless, A. C., & Lawrence, J. G. (2010). Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(25), 11453–11458. http://dx.doi.org/10.1073/pnas.1001291107.

Rokas, A. (2001). Phylogenetic analysis of protein sequence data using the Randomized Axelerated Maximum Likelihood (RAXML) program. In F. M. Ausubel & R. E. Kingston, et al. (Eds.), *Current protocols in molecular biology: 96:19.11.1-14*. Hoboken, NJ, USA: Wiley. Retrieved from, http://onlinelibrary.wiley.com/doi/10.1002/0471142727.mb1911s96/abstract.

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, *425*(6960), 798–804. http://dx.doi.org/10.1038/nature02053.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, *4*(1), 41. http://dx.doi.org/10.1186/1471-2105-4-41.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090. http://dx.doi.org/10.1073/pnas.74.11.5088.

Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., & Kellis, M. (2014). Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Research*, *24*(3), 475–486. http, //dx.doi.org/10.1101/gr.161968.113.

Yu, Y., Barnett, R. M., & Nakhleh, L. (2013). Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology*, *62*(5), 738–751. http://dx.doi.org/10.1093/sysbio/syt037.