Inference of Gain and Loss Events from Phyletic Patterns Using Stochastic Mapping and Maximum Parsimony—A Simulation Study

Ofir Cohen[1] and Tal Pupko[1,2,*]

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[2]National Evolutionary Synthesis Center, Durham, North Carolina

*Corresponding author: E-mail: talp@post.tau.ac.il.

Accepted: 27 September 2011

Abstract

Bacterial evolution is characterized by frequent gain and loss events of gene families. These events can be inferred from phyletic pattern data—a compact representation of gene family repertoire across multiple genomes. The maximum parsimony paradigm is a classical and prevalent approach for the detection of gene family gains and losses mapped on specific branches. We and others have previously developed probabilistic models that aim to account for the gain and loss stochastic dynamics. These models are a critical component of a methodology termed stochastic mapping, in which probabilities and expectations of gain and loss events are estimated for each branch of an underlying phylogenetic tree. In this work, we present a phyletic pattern simulator in which the gain and loss dynamics are assumed to follow a continuous-time Markov chain along the tree. Various models and options are implemented to make the simulation software useful for a large number of studies in which binary (presence/absence) data are analyzed. Using this simulation software, we compared the ability of the maximum parsimony and the stochastic mapping approaches to accurately detect gain and loss events along the tree. Our simulations cover a large array of evolutionary scenarios in terms of the propensities for gene family gains and losses and the variability of these propensities among gene families. Although in all simulation schemes, both methods obtain relatively low levels of false positive rates, stochastic mapping outperforms maximum parsimony in terms of true positive rates. We further studied the factors that influence the performance of both methods. We find, for example, that the accuracy of maximum parsimony inference is substantially reduced when the goal is to map gain and loss events along internal branches of the phylogenetic tree. Furthermore, the accuracy of stochastic mapping is reduced with smaller data sets (limited number of gene families) due to unreliable estimation of branch lengths. Our simulator and simulation results are additionally relevant for the analysis of other types of binary-coded data, such as the existence of homologues restriction sites, gaps, and introns, to name a few. Both the simulation software and the inference methodology are freely available at a user-friendly server: http://gloome.tau.ac.il/.

Key words: phyletic pattern, stochastic mapping, maximum parsimony, evolutionary models.

Introduction

Gene Content Modifications among Microbial Species

Evolutionary biologists had long recognized that gain and loss of genetic material are a central mechanism augmenting site-specific mutations in the evolution of microbial species (Achtman and Wagner 2008). Recent advances in genome sequencing elucidate the extent in which these macro evolutionary events are responsible for microbial genome remodeling (Konstantinidis and Tiedje 2004; Koonin and Wolf 2008). Modifications in microbial gene content are pivotal in the adaptation to new environments. Examples include genome erosions that facilitate endosymbiosis (Moran et al. 2009), the acquisition of novel genes that are associated with adaptation to new ecological niches (Gogarten and Townsend 2005), attainment of novel functions (Pennisi 2004; Gogarten and Townsend 2005), expansion of metabolic networks (Pal et al. 2005), speciation (Lawrence 1999), and pathogenicity transformation (Jin et al. 2002; Holden et al. 2004; Gal-Mor and Finlay 2006).

Three approaches are typically used to infer gene transfer events, each suitable for inferring only a subset of all transfer events. The so-called "phylogenetic incongruence" approach identified genes with incompatible evolutionary history as compared with the inferred ribosomal trees (Sicheritz-Ponten and Andersson 2001). This approach is suitable for relatively widespread genes with "not too much or too little" sequence divergence (e.g., Graybeal 1994). The so-called parametric genomic composition approach detects genes that are significantly different from the rest of the genome in some attributes such as GC content or codon usage (Lawrence and Ochman 1998). This approach can only detect recent transfer events due to sequence amelioration (Koski et al. 2001; Wang 2001; Daubin et al. 2003). Finally, phyletic pattern–based approaches rely on the availability of fully sequenced genomes and can capture the emergence of new gene family on a background of their absence in closely related species.

Phyletic Patterns and Detection of Gain and Loss Events

Comparative genomic analysis of gene gain and loss events across multiple species requires compact representation of gene content. A set of genomes is represented by a matrix of binary characters that resembles a gap-free multiple sequence alignment and is often termed a phyletic pattern or a phylogenetic profile. Rows correspond to species and columns to gene families. The character in row *i* and column *j* is either "1" or "0" depending on whether gene family *j* is present or absent in species *i*, respectively.

Such a binary presence–absence matrix is used to represent numerous other biological data including restriction sites (Templeton 1983; Nei and Tajima 1985; Felsenstein 1992), indels (Simmons and Ochoterena 2000), introns (Csuros 2006; Carmel et al. 2007), and morphological characters (reviewed in Ronquist 2004). Notably, even questions in fields other than biology are amenable to such data coding. For example, the evolution of human languages was studied by analyzing the phyletic patterns of lexical units (Gray and Atkinson 2003).

Following the development of realistic probabilistic models describing the evolution of DNA and protein sequences, the analysis of phyletic pattern data has progressed from the maximum parsimony criterion (Mirkin et al. 2003; Boussau et al. 2004) to models, in which the dynamics of gain ($0 \rightarrow 1$) and loss ($1 \rightarrow 0$) events are assumed to follow a continuous-time Markov process (Csuros 2006; Hao and Golding 2006). Recent advances in probability models for analyzing gene content data allow more realistic description of the evolutionary dynamics of gene family gains and losses. For example, a recent model by Spencer and Sangaralingam (2009) allows variability of the gain and loss rates among branches to be explicitly accounted for. This model improvement is important for analyzing gene content changes in lineage leading to parasitic bacteria, in which massive gene losses are often observed (Moran 2003; Charlebois and Doolittle 2004; Moran et al. 2009). In another example, variability of both the gain and the loss rates is allowed among gene families, thus alleviating the unrealistic assumption that all gene families evolve with a single gain–loss ratio (Cohen and Pupko 2010).

One of the goals when analyzing phyletic pattern data is to map gain and loss events onto a phylogenetic tree. In gene family analysis, this corresponds to inferring for each gene family the branches in which this gene family was acquired (gained the first copy of the gene) or lost (all copies deleted). A prevailing branch-site detection methodology is based on the maximum parsimony approach. Parsimony-based mapping is used in many recent works (Kettler et al. 2007; Cordero et al. 2008; Lercher and Pal 2008; Ruano-Rubio et al. 2009; Yerrapragada et al. 2009; Kloesges et al. 2011). However, an alternative methodology exists, in which evolutionary events are mapped onto the phylogeny within a probabilistic paradigm (Nielsen 2002; Huelsenbeck et al. 2003; Bollback 2005; Minin and Suchard 2008). This stochastic mapping methodology allows exact computation of both the expectation and the probability of transitions along each branch of a phylogenetic tree, given the evolutionary models and the phyletic pattern data. All possible ancestral paths are accounted for, weighted by their likelihood (Cohen and Pupko 2010).

It has been shown in several cases that using the maximum parsimony criterion for sequence analysis may be misleading, in particular when there is substantial variability in branch lengths (Felsenstein 1978; Yang 1996; Pol and Siddall 2001; Swofford et al. 2001). However, although the developers of the stochastic mapping approach had performed initial performance evaluation (Nielsen 2002; Huelsenbeck et al. 2003), a rigorous comparison between maximum parsimony and the stochastic mapping for the task of inferring gain and loss events to specific branches is still missing. Here, we evaluated the performance of these two approaches for detecting branch-site gain and loss events under various evolutionary assumptions. We further aimed to study the parameters that determine the inference accuracy of each methodology. For this performance evaluation, we have developed a simulation program, which allows simulating phyletic pattern data under various scenarios of gain and loss dynamics.

Materials and Methods

Simulations

The simulation software is given an underlying phylogeny that represents the species tree and a set of assumptions regarding the evolutionary dynamics of gain and loss events, parameterized as a continuous-time Markov chain. In this

analysis, for all sites, evolution is simulated along the same tree. During simulations, all gain and loss events along each branch for each site (gene family) are recorded.

In all simulations, stationarity was assumed, thus character frequencies at the root were set to the stationary frequencies of the rate matrix. The rate matrix, sampled for each site (gene family), governs the evolutionary dynamics for this site and thus determines substitution probabilities along the tree. The total rate of a specific matrix is defined as the stationary frequency of 1 (π_1) times the gain rate plus the stationary frequency of 0 (π_0) times the loss rate. All matrices were scaled so that the average total rate over all simulated sites equals 1. This ensures that the branch lengths used in the simulated tree correspond to average number of gain and loss events per site. Simulations were conducted under several evolutionary scenarios starting with a naïve scenario with equal gain and loss rates and no rate variability among different sites (ER_gEql).

Simulations with Variable Loss-to-Gain Ratio

The assumption that gain and loss rates are equal in all sites is alleviated by sampling for each site the loss–gain rate ratio from a uniform distribution. We simulated several variants, in which we progressively introduced a bias toward a higher loss-to-gain ratio. Specifically, the loss-to-gain rate ratio was sampled from a uniform distribution in the interval [0, 2 × expectedRatio]. Thus, when expectedRatio = 1, the loss-to-gain ratio was sampled from the interval [0, 2], and the expectation of the ratio is 1. We denote this simulation scenario as ER_gVrl_1, in which the suffix number stands for the expectation of the loss-to-gain ratio. Similarly, we simulated scenarios ER_gVrl_2, ER_gVrl_4, and ER_gVrl_8. To avoid boundary conditions, ratios were sampled uniformly from the interval $[\epsilon, 2 \times ratio - \epsilon]$, with ϵ set to 0.01. For each site, we derived the gain and loss rates while maintaining the overall rate for that site equal to 1.

Simulations with Rate Variability among Sites

Additional scenarios further alleviated the assumption that all sites evolve under the same total rate. The rate variability among sites was implemented by sampling from a gamma distribution, which was shown to capture well the rate variability in gain and loss dynamics among gene families (Cohen et al. 2008; Hao and Golding 2008b). All previous scenarios that assume a single rate for all sites were modified to account for among sites rate variability (with name prefix changed from "ER" to "VR"). The rate variability may be considered a "second layer" of variability in our implementation. We thus sampled two variables for each site: the loss-to-gain rate ratio (as before) and the overall evolutionary rate. For all simulations, we set the shape parameter of the gamma distribution to 0.6, which is suited for the rate variability found in gene families across microbial species (Cohen et al. 2008; Hao and Golding 2008b; Spencer and Sangaralingam 2009).

Simulations of Evolutionary Dynamics Derived from COG Gene Families

We also simulated data with gain and loss dynamics based on real data: phyletic pattern data including 4,873 gene families across 66 microbial genomes extracted from the Clusters of Orthologous Groups (COG) database (Tatusov et al. 2003) using the underlying phylogeny from the "Tree Of Life" project (Ciccarelli et al. 2006). Based on this data set, two related simulation scenarios were established. In simulation scenario $COG_{Parsimony}$, maximum parsimony inference was used to infer the evolutionary parameters (gene families' rate distributions) in the simulations while using a cost matrix (gain:loss) of 2:1 (Snel et al. 2002). This distribution was computed as follows: for each gene family, the gain and loss rates were proportional to the number of gain and loss events inferred for that gene family, respectively. Simulations were then conducted by sampling for each simulated site, a (gain, loss) pair from the COG gene families with repetition. In COG_{Model}, evolutionary rates were based on a COG-fitted evolutionary model. Specifically, a gain–loss mixture model was assumed, and the model parameters were estimated using maximum likelihood (ML) from the COG gene family data (Cohen and Pupko 2010). The estimated parameters determine two gamma distributions, one for the gain rate parameter and one for the loss rate parameter (see supplementary table S2, Supplementary Material online). Simulations were then conducted in COG_{Model} scenario by sampling gain and loss rates from the gain and loss gamma distributions obtained empirically.

Inference Methods

In the consecutive step, the resulting simulated phyletic pattern and the species tree (only topology) are given as input for both the maximum parsimony and the stochastic mapping methods, which infer gain and loss events for each gene family and for each branch. The stochastic mapping method assumes an evolutionary model. Here, we used a stationary model allowing variability among genes for both gain and loss rates (Cohen and Pupko 2010). The model's free parameters and phylogeny branch lengths are unknown and are estimated numerically based on the simulated phyletic pattern using the ML criterion. Maximum parsimony events detection is based on the Sankoff reconstruction method with adaptable cost matrix (Sankoff 1975).

Performance Evaluation

Performance of both methods is evaluated by considering gain and loss inference as a binary classification problem. For each branch and site, the method has to correctly predict whether a gain event has occurred or not and similarly for loss events.

Comparable Recalls Ratio Measure of Performance

Event detection by stochastic mapping is determined by a varying cutoff value (posterior probability of event), thus multiple classifications are possible with various levels of sensitivity (=true positive rate [TPR]) and specificity (=1 − false positive rate [FPR]). However, maximum parsimony detection results with a single classification. Thus, instead of comparing Receiver Operating Characteristic (ROC) curves and the consequent Area Under the Curve (AUC), we used comparable recall—the sensitivity (recall) of both methodologies while maintaining the same specificity. To compute comparable recalls, the recalls of maximum parsimony and stochastic mapping must be measured with the same FPR. Thus, the recall of stochastic mapping was measured with a cutoff that corresponds to an FPR, which is equal to (or slightly lower than) that of the maximum parsimony approach. In practice, given the finite number of cutoff values, it was impractical to set the FPR of the stochastic mapping approach to be identical to that of maximum parsimony. Thus, the cutoffs used for stochastic mapping (posterior probability for events occurrence) were chosen to be conservative, that is, the FPR of the stochastic mapping was always the highest possible cutoff that is still lower than that of maximum parsimony.

Matthews Correlation Coefficient Measure of Performance

The Matthews Correlation Coefficient (MCC) is a relatively balanced measure of classification performance. MCC values vary between −1 and +1 and are interpreted as the correlation between the set of predictions and the set of simulations (Matthews 1975; Baldi et al. 2000). MCC computations use all four numbers: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

Data Set Size in Simulations

For each scenario, we simulated 10,000 sites independently, assembled into a phyletic pattern used for the inference by both methods. The underlying tree used in these simulations contains 66 species. Because inference is performed for each site and each branch, the sample size in each simulation equals the number of sites multiplied by the number of branches, equals 1,300,000 for either gain or loss inference and 2,600,000 when overall performance for both events is considered.

Analyzing Factors That Determine Performance

When analyzing the evolutionary rate as the factor that determines performance, we used the $\mathrm{COG}_{\mathrm{Parsimony}}$ simulation scenario and evaluated inference restricted to subsets of gene families according to the total evolutionary rate.The total rate of a specific matrix, as defined above, was used to classify sites as either fast (rate higher than 1) or slow (all the other sites) evolving.

When analyzing the number of sites in the data set as a factor determining performance, we have performed inference with only a limited number of sites (n). In this case, the model parameters and branch lengths that are estimated by stochastic mapping were based on n sites. The total number of sites for performance analysis was 10,000 in all cases. Thus, we performed 10,000/n-replicated simulations of the same scenario.

Results

Our main interest is to evaluate the performance of stochastic mapping and maximum parsimony in accurately detecting lineage-specific gain and loss events along a phylogenetic tree. Inference of branch-site specific gain or loss events is formalized as a classification task. Specifically, we used two procedures to estimate detection performance. The first comparative performance procedure measures the different levels of sensitivity (recall or TPR) of both methodologies while maintaining the same specificity (complement of FPR). The FPR used in the comparison is determined by the maximum parsimony method (i.e., the stochastic mapping detection cutoff is set to match the maximum parsimony's FPR). We term this value Comparable Recalls Ratio (CRR, details in Materials and Methods). The second evaluation procedure employs the MCC, in which values of 1 and 0 represent perfect correlation between simulated and inferred events and random prediction capacity, respectively (Matthews 1975; Baldi et al. 2000). We used these two measures to gain insights regarding factors determining the accuracy of both mapping methodologies.

Performance under Various Simulation Scenarios

We start by simulating phyletic pattern data under a naïve evolutionary model, in which all sites (e.g., gene families) evolve with the same rate, and gain and loss rates are equal to each other. In this simulation scheme, maximum parsimony obtained overall FPR of 0.006 and TPR (recall) of 0.421. This detection rate is also evaluated with MCC value of 0.563 (Matthews 1975). Given the same FPR, stochastic mapping obtains TPR of 0.763 and MCC value of 0.809. Thus, recall and MCC values for stochastic mapping were 81.2% and 43.5% higher compared with maximum parsimony given the same FPR, respectively. These results are depicted in Table 1 (simulation scenario code name ER_gEql).

Performance under Variable Loss-to-Gain Ratio Variability

The simplified assumptions in the above evolutionary scenario were relaxed in subsequent scenarios. We alleviated the assumption that all sites evolve with the same gain and loss rates and that the loss-to-gain ratio equals 1 for

Table 1
Evaluation of Stochastic Mapping and Maximum Parsimony Performance for Events Detection in Various Simulation Schemes

Simulation Scenario Code	Rate Distribution among Sites	Loss/Gain Ratio in Simulation	CRR	MCCs Ratio	MCC Mapping	TPR Mapping	FPR Mapping	MCC Parsimony	TPR Parsimony	FPR Parsimony
ER_gEql	Equal	1[b]	1.812	1.435	0.809	0.763	0.005	0.563	0.421	0.006
ER_gVrl_1	Equal	1	1.515	1.296	0.685	0.577	0.005	0.529	0.381	0.006
ER_gVrl_2	Equal	2	1.588	1.345	0.666	0.563	0.006	0.495	0.354	0.006
ER_gVrl_4	Equal	4	1.709	1.426	0.616	0.503	0.007	0.432	0.294	0.007
ER_gVrl_8	Equal	8	1.675	1.416	0.524	0.381	0.006	0.37	0.227	0.006
VR_gEql	Gamma	1[b]	1.834	1.463	0.729	0.651	0.005	0.498	0.355	0.006
VR_gVrl_1	Gamma	1	1.952	1.527	0.712	0.621	0.005	0.466	0.318	0.005
VR_gVrl_2	Gamma	2	2.007	1.557	0.702	0.608	0.005	0.451	0.303	0.005
VR_gVrl_4	Gamma	4	2.093	1.608	0.66	0.546	0.005	0.411	0.261	0.005
VR_gVrl_8	Gamma	8	2.142	1.636	0.593	0.446	0.004	0.362	0.208	0.004
COG_Parsimonyy	Parsimony[a]	2.89	1.359	1.22	0.425	0.246	0.004	0.348	0.181	0.004
COG_Model	Model[a]	4.63	1.802	1.456	0.576	0.419	0.004	0.396	0.232	0.004

[a] Rates based on empirical estimation of COG gene families.
[b] The gain-to-loss ratio is 1 and does not vary among sites.

all sites. Instead, we allowed the loss-to-gain rate ratio to vary among sites. To model this variation, we simulated under the assumption that the loss-to-gain rate ratio is distributed uniformly between 0 and 2, so that the loss-to-gain rate ratio has equal probabilities to be higher or lower than 1 (simulation scenario code name ER_gVrl_1, Table 1). In subsequent simulations, we sampled the loss-to-gain ratio from uniform distributions between 0 and either 4, 8, or 16, thus biasing the simulations to increasingly higher loss rates (details in Materials and Methods). The increased expected loss-to-gain ratio is designated by median loss-to-gain ratio of 2, 4, and 8 (simulation scenario code names ER_gVrl_r, where "r" designates the median ratio).

Using the MCC to evaluate inference performance, which allows comparison across different simulations, we compare the performance given the increase in loss-to-gain rates ratio. Our results indicate that increased loss-to-gain ratio reduces the accuracy of both inference methods as MCC values monotonically decrease as the loss-to-gain ratio increases. To illustrate this trend, the MCC values dropped from 0.529 to 0.37 and from 0.685 to 0.524, for maximum parsimony and stochastic mapping, respectively, when comparing loss-to-gain ratio of 1 to loss-to-gain ratio of 8 (compare ER_gVrl_1 with ER_gVrl_8, Table 1). These results indicate that variable loss-to-gain ratio decreases accuracy for both methods. Notably, in all of these simulation scenarios, stochastic mapping had significantly higher performance than maximum parsimony. However, there is no consistent trend in the relative performances of both methods (measures in terms of CRR and MCC ratio) when the loss-to-gain ratio increases.

Performance under Rate Variability among Sites

We further alleviated the assumption that all sites evolve under the same total gain + lost rate. We repeated the above-mentioned five scenarios, but in these simulations, gain + loss rates variability among sites is allowed (details in Materials and Methods). When the total gain + loss rate varies among sites, inference is more difficult for maximum parsimony but not for stochastic mapping (compare code name starting with ER with those starting with VR with the same loss-to-gain ratio; Table 1). For example, the MCC maximum parsimony score dropped from 0.529 to 0.451 when the total rate was allowed to vary (ER_gVrl_1 vs. VR_gVrl_1). In contrast, for stochastic mapping, the MCC increased from 0.685 to 0.712. Thus, in all cases, the performance difference between stochastic mapping and maximum parsimony became higher when rate variability is allowed. The higher performance of stochastic mapping over maximum parsimony is most pronounced when comparing equal rates versus variable rates for higher loss-to-gain ratios (Table 1, ER_gVrl_8 vs. VR_gVrl_8). CRR and MCC ratios are 1.675 and 1.416 for equal rates and 2.142 and 1.636 for variable rates, respectively. These results suggest that maximum parsimony is sensitive to variation both in the total rate and in the loss-to-gain ratio, which together contribute to the relative poor performance of maximum parsimony.

Simulation with Empirical Distributions of Gene Families Dynamics

The above evaluation of performance is based on simulations in which gain and loss rate distributions across sites are based on theoretical distributions. Although in the more complex simulation schemes above, we allowed both the loss-to-gain rate ratio and the overall rate to vary among sites, these are still oversimplified scenarios that may poorly represent real evolutionary histories. Aiming for more realistic evolutionary simulations in terms of gain and loss dynamics, we simulated phyletic patterns with gain and loss rates that were estimated from a real data set of microbial

species' phyletic pattern (based on COG gene families; details in Materials and Methods).

Two simulation scenarios were used. In the first ($COG_{Parsimony}$), the empirical gene family dynamics were estimated by maximum parsimony, whereas in the second (COG_{Model}), it was estimated using an evolutionary model with parameters fitted to the data. Intuitively, $COG_{Parsimony}$ simulations may favor parsimony-based inference. Indeed in this simulation scenario, the difference in performance was smaller than all previous simulation scenarios. Nevertheless, stochastic mapping performance was still significantly higher than maximum parsimony. In this simulation, CRR indicates 35.9% higher comparable TPR for stochastic mapping–based inference over maximum parsimony-based inference. The comparable MCC is also 22% higher. Under COG_{Model}, as expected, the comparable recalls and MCC performance values were 80.2% (CRR) and 45.6% (MCC) higher for stochastic mapping (Table 1).

Maximum Parsimony Cost Matrix

The results presented above were based on a naïve maximum parsimony inference in which gain and loss events were given equal costs. Several studies analyzing the evolution of gene families modified the cost matrix by assuming that the cost of gain events is double that of loss events (Snel et al. 2002; Pal et al. 2005). We repeated the analysis above for the simulations based on rates estimated by maximum parsimony from COG gene families ($COG_{Parsimony}$), this time comparing the performance of maximum parsimony with a cost matrix of 1:1 versus a cost matrix of 2:1 (i.e., the cost of a gain event is twice that of a loss event). We observe a relatively small difference in maximum parsimony performance with gain cost double that of loss: the MCC values were 0.348 and 0.337 for costs of 1:1 and 2:1, respectively (Table 2).

Inference under higher costs for gain events should result in more conservative inference of gain events and vice versa for loss events (i.e., equivalent to higher gain inference threshold and lower for loss). To test this expectation, we repeated performance evaluation separately for gain detection and for loss detection. As expected, when the cost of gain events is raised, the maximum parsimony number of gain events inferred is decreased as evident with lower TPR from 0.231 to 0.163 and lowered FPR from 0.003 to 0.001. The opposite trend is observed with loss detection such that the increased gain cost results with more loss events detected and thus higher TPR and FPR in loss performance evaluation (Table 2). The trend exemplified in these scenarios was observed in all simulation scenarios (supplementary table S1, Supplementary Material online), namely that modifying the cost matrix to reflect the simulation scenario, loss-to-gain ratio does not necessarily improve the overall performance of maximum parsimony. When the cost matrix is adjusted to account for higher simulated loss propensity (by higher gain cost), the sensitivity for loss detection is increased, whereas the sensitivity for gain detection is decreased.

Table 2
Maximum Parsimony Performance Separated for Gain and Loss Detection under Two Parsimony Cost Matrices

	Cost Matrix (Gain:Loss)	MCC	TPR	FPR
Overall inference	Cost 1:1	0.348	0.181	0.004
	Cost 2:1	0.337	0.18	0.004
Gain inference	Cost 1:1	0.388	0.231	0.003
	Cost 2:1	0.356	0.163	0.001
Loss inference	Cost 1:1	0.322	0.131	0.001
	Cost 2:1	0.339	0.197	0.003

NOTE.—The simulation scenario in all these evaluations is based on rates estimated by maximum parsimony from COG gene families ($COG_{Parsimony}$).

Detailed Performance Evaluation—Factors Determining Performance

The results presented so far were averaged over all simulated sites and branches. In this section, we study several parameters that determine performance and reevaluate the two methods with respect to these parameters. These parameters include 1) internal versus external branches, 2) fast versus slowly evolving sites, and 3) the effect of data set size. In all the results presented below, simulations were conducted under the $COG_{Parsimony}$ scenario and maximum parsimony inference was based on equal gain and loss costs.

External versus Internal Branches

We compared the inference restricted with specific subsets of branches. We inferred events either along external branches (those leading to an extant species in the tree), along internal branches (those not leading to an extant species in the tree), or deep branches (branches that neither lead to extant species nor to direct father of extant species). In Table 3, we provide the performance of stochastic mapping and maximum parsimony for each of these subsets as well as for all branches (called "Reference" in Table 3). Results for all other simulated scenarios are provided in supplementary table S3 (Supplementary Material online). As expected, our results indicate that for both methodologies, inference along external branches is more accurate compared with overall branches, which in turn is more accurate than the inference along internal branches. Inference along deep branches is even less accurate. However, the performance of maximum parsimony was more substantially reduced when moving toward internal branches as compared with stochastic mapping: performance ratios between stochastic mapping and maximum parsimony (CRR) increased from 1.208 (external branches) through 1.562 (internal branches) to 2.044 (deep branches). Similar results are observed when MCC ratios are compared (Table 3).

Table 3
Performance Evaluation in Various Subsets of Events

Evaluated Subset	CRR	MCCs Ratio	MCC Mapping	TPR Mapping	FPR Mapping	MCC Parsimony	TPR Parsimony	FPR Parsimony
Reference	1.359	1.22	0.425	0.246	0.004	0.348	0.181	0.004
External branches	1.208	1.125	0.543	0.378	0.003	0.483	0.313	0.003
Internal branches	1.562	1.352	0.357	0.185	0.004	0.264	0.119	0.004
Deep branches	2.044	1.795	0.242	0.11	0.005	0.135	0.054	0.005
Low rate	1.208	1.123	0.494	0.309	0.002	0.44	0.256	0.002
High rate	1.47	1.301	0.387	0.217	0.005	0.297	0.147	0.005

NOTE.—The simulation scenario in all these evaluations is based on rates estimated by maximum parsimony from COG gene families ($COG_{Parsimony}$).

We explain these results by the observation that inference of evolutionary events requires reconstruction of ancestral states. In many cases, reconstruction of ancestral states is more error prone at deeper nodes of the trees as the distance from the known states at the leaves increases. In other words, uncertainty in ancestral states reconstruction may pose a greater challenge to the maximum parsimony method than the stochastic mapping method, resulting with increased error rates.

Fast versus Slow Evolving Gene Families

We compared the performance of both methods for fast versus slow evolving gene families (see Materials and Methods). Table 3 lists performance for each of these subsets as well as for all gene families (called Reference in Table 3).

Our results indicate that higher underlying rate results with lower performance. Comparing performance evaluated for the low-rate group with that of the high-rate group, the MCC values decreased from 0.44 to 0.297 and from 0.494 to 0.387 for maximum parsimony and stochastic mapping, respectively. Importantly, accurate inference of events occurring within gene families with higher evolutionary rate is more difficult for maximum parsimony than to stochastic mapping. Thus, the comparative performance ratios increased between the low- and the high-rate groups —from 1.208 to 1.47 and from 1.123 to 1.301 for CRR and MCCs ratio, respectively. These results may be explained by the fundamental parsimonious principle—minimizing the number of events (weighted by their cost). Thus, with simulations under higher mean rate, there is higher probability for a violation of the parsimonious principle with evolutionary history that includes multiple events. Although such evolutionary history is harder to reconstruct regardless of the method used, our results indicate that stochastic mapping inference is more robust with respect to higher evolutionary rates.

Evaluating Performance with Variable Data Set Size

Here, we test one fundamental difference between stochastic mapping and maximum parsimony inference methodologies. Maximum parsimony method is model free, whereas stochastic mapping is based on an underlying evolutionary model. Notably, maximum parsimony inference is conducted under a specific cost matrix, but these costs are assumed rather than evaluated from the data. The evolutionary model and branch lengths are estimated from all simulated sites as the first step of stochastic mapping inference (details in Materials and Methods).

We repeated the $COG_{Parsimony}$ simulation scenario but instead of allowing stochastic mapping to use 10,000 simulated sites to estimate the model parameters and branch lengths, we replicated the simulation scenario, each time with a smaller number of sites (for details, see Materials and Methods). The results summarized in Table 4 depict performance evaluation with variable number of sites used for model estimation. Although maximum parsimony performance did not vary as a function of the number of sites, stochastic mapping performance monotonically decreases with smaller number of sites (Table 4). When the number of sites was reduced from 10,000 to 10, comparative ratios decreased dramatically—from 1.359 to 0.898 and from 1.22 to 0.941 for CRR and MCCs ratio, respectively. Thus, when only 10 sites were available for the evaluation of stochastic mapping's required parameters, maximum parsimony inference was more accurate. Interestingly, our results indicate that with as few as 50 sites, stochastic mapping performance surpasses that of maximum parsimony by 16.6% and 10.6% for CRR and MCCs ratio, respectively.

A further simulation scheme reveals that the high error rates by stochastic mapping with limited number of sites is due to unreliable branch length estimation rather than the evolutionary model parameters. When the input data were limited to 10, and the "true" branch lengths were provided rather than estimated from the data, a remarkably high performance was observed for stochastic mapping: 2.43 and 1.74 for CRR and MCCs ratio, respectively. Taken together, these results suggest that small data set substantially reduce the performance of stochastic mapping. However, these results also suggest that stochastic mapping inference is highly robust to model parameter misspecification, when branch lengths are given.

Evaluating Performance Reproducibility

Here, we evaluated the reproducibility of stochastic mapping performance. Because stochastic mapping requires model parameters and branch length estimation, stochastic mapping performance varies in each simulation depending

Table 4
Performance Evaluation with Variable Data Set Size

Number of Sites Used for Model and Branch Lengths Estimation	CRR	MCCs Ratio	MCC Mapping	TPR Mapping	FPR Mapping	MCC Parsimony	TPR Parsimony	FPR Parsimony
10,000	1.359	1.22	0.425	0.246	0.004	0.348	0.181	0.004
5,000	1.34	1.21	0.425	0.247	0.00354	0.352	0.184	0.00355
1,000	1.34	1.21	0.423	0.245	0.00352	0.35	0.183	0.00357
500	1.32	1.2	0.421	0.243	0.00351	0.351	0.183	0.00352
100	1.23	1.15	0.4	0.224	0.00354	0.349	0.182	0.00357
50	1.18	1.12	0.394	0.219	0.0035	0.353	0.185	0.00351
10	0.898	0.941	0.328	0.163	0.00328	0.349	0.181	0.00351
10[a]	2.43	1.74	0.604	0.44	0.00357	0.348	0.181	0.00359

NOTE.—Smaller number of sites available for model and branch length estimation results with lowered stochastic mapping performance. The simulation scenario in all these evaluations is based on rates estimated by maximum parsimony from COG gene families ($COG_{Parsimony}$). In all cases, overall number of sites used for performance estimation was 10,000.

[a] Branch lengths are given rather than estimated from the data.

on the accuracy of these parameters. The $COG_{Parsimony}$ simulation scenario was replicated 20 times, each replication with 1,000 simulated sites. The performance of each replication was analyzed separately. Stochastic mapping performance was higher than that of maximum parsimony in all 20 replications. Remarkable reproducibility was observed indicated with highly similar comparative ratios among replications. Average values were 1.34 and 1.21, standard errors (SEs) were 0.008 and 0.005, and minimal values were 1.29 and 1.18, for CRR and MCCs ratio, respectively (for both CRR and MCCs ratio, P value $< 10^{-100}$, Z-test).

The $COG_{Parsimony}$ is the simulation scenario in which the difference in performance between stochastic mapping and maximum parsimony is the smallest. Thus, the differences between stochastic mapping and maximum parsimony for all other scenarios (with 10,000 sites) are also highly statistically significant (data not shown).

Running Times

Occasionally, users may favor a fast methodology over a more accurate one, which is computationally intensive and thus requires longer running times. We compared running times required for gain and loss inference for both methods. Running times of the stochastic mapping approach depend on the number of discrete categories assumed in the gain–loss mixture model. Here, we used four discrete categories for gain events and four discrete categories for loss events (Cohen and Pupko 2010). To compute running times, both methods inferred events for 1,000 sites along 130 branches. Computations were conducted using an AMD Opteron Processor 2356 at 2.2 GHz. As expected, the maximum parsimony method was substantially faster, taking on average 0.023 min (SE = 0.0017) compared with 9.56 min (SE = 0.42) for the entire stochastic mapping procedure. Notably, although maximum parsimony is a much faster method, this analysis shows that stochastic mapping inference can be obtained in a couple of minutes for data sets of ordinary size.

Discussion

Recently, parsimony-based methods to analyze phyletic pattern were augmented by several probabilistic evolutionary models. The stochastic mapping method, based on such models, allows explicit quantification of the probability and expectation for gain and loss events for each site and branch. In this study, we performed extensive evaluations of the ability of the maximum parsimony and the stochastic mapping approaches to accurately map such lineage-specific events. Our simulation-based results reveal various factors that determine inference accuracy by both methodologies. We have used two comparative measurements for performance accuracy—comparing recall rates given the same FPR (termed CRR) and MCCs ratio (Matthews 1975). These comparative values revealed simulation schemes and factors resulting with higher or smaller differences between these two methods. However, the emerging conclusion is that in all but one case, stochastic mapping performance is significantly higher.

Arguably, the higher performance by the probabilistic stochastic mapping approach is expected, as it was often demonstrated that ML outperforms maximum parsimony in phylogeny and ancestral state reconstructions (Felsenstein 1978; Yang 1996; Pol and Siddall 2001; Swofford et al. 2001). However, our goal here was to rigorously study the performance of both methods, focusing on phyletic patterns analysis and the specific parameters that determine gain and loss detection accuracy. For example, in phyletic pattern data, the evolutionary dynamics depend on the gain–loss rate ratio and the total rate variability. We demonstrated that the

maximum parsimony performance varies substantially depending on these factors, although stochastic mapping performs well for a large set of scenarios. Additionally, we found substantial accuracy reduction in detection of ancient events (occurring along deep branches) and in detection of events for fast evolving gene families (i.e., governed by fast gain and loss rates). Our analyses also reveal that in these more challenging conditions, maximum parsimony error rate becomes substantially higher than stochastic mapping. We additionally illustrate the dependence of accurate stochastic mapping on the number of sites in the phyletic data. We find that for very small data sets, the expected error of stochastic mapping is considerably large. Taken together, our study allows better understanding of the factors that determine the inference accuracy of both methods.

Inference, based on phyletic patterns, and our simulations have a few limitations. First, our simulation study most likely overestimates accuracy levels for both methods. Main factors that are expected to reduce accuracy and are ignored here include missed organisms by sampling and extinctions (e.g., Heath et al. 2008), uncertainty in reconstruction of the phylogenetic tree for the extant species (Ronquist 2004), and inaccurate classification of gene families (Zhaxybayeva et al. 2007; Hao and Golding 2008a). However, there are no indications that these factors differentially influence the two methodologies. Our results indicate that stochastic mapping performance is highly dependent on reliable branch length estimation. A Bayesian approach for phyletic pattern analysis that takes into account uncertainty in branch lengths can alleviate this sensitivity.

Our study focuses on the inference of gain and loss events of gene families during the evolution of microbial species. This presence-absence–based analysis is biologically justified as it captures major changes in the proteome composition of the host genomes (Pal et al. 2005). Nevertheless, in such a phyletic pattern representation of the data, the number of paralogs for each gene family is ignored and duplications or reductions in the number of paralogs cannot be detected. Clearly, a richer Markovian model accounting for the number of genes within each gene family will better capture gene family dynamics. Notably, a projection of such a richer model onto a binary alphabet would result in a non-Markovian behavior (i.e., pulling together all copy numbers greater than zero into a single state 1 makes the process non-Markovian). This argument suggests that our simulation settings, in which a Markovian process is assumed both for the simulations and the stochastic mapping inference, may overestimate the performance of stochastic mapping–based inference. Clearly, further work is needed to extend the stochastic mapping approach to analyze the evolution of the number of paralogs along the tree. Nonetheless, there are many cases in which the usage of phyletic patterns is not a compact representation of copy number variation. To this end, our phyletic pattern simulations results are also valuable for binary data such as restriction sites (Felsenstein 1992), indels (Simmons and Ochoterena 2000), introns (Csuros 2006; Carmel et al. 2007), morphological characters (Ronquist 2004), and even gain and losses of lexical units (Gray and Atkinson 2003).

Maximum parsimony is still a widely used approach for analyzing phyletic data (Pal et al. 2005; Kettler et al. 2007; Cordero et al. 2008; Lercher and Pal 2008; Ruano-Rubio et al. 2009; Yerrapragada et al. 2009; Georgiades et al. 2011; Kloesges et al. 2011). Our study shows that branch-specific gain and loss events inference is more accurate with the probabilistic stochastic mapping method compared with maximum parsimony, for a wide range of evolutionary scenario. The complete phyletic pattern analysis methodology and the simulation software are freely available in a user-friendly web server (http://gloome.tau.ac.il/; Cohen et al. 2010).

Supplementary Material

Supplementary tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Haim Ashkenazy for his help with the web server implementation. We thank David Burstein for critically reading the manuscript. We also thank Associate Editor David Bryant, Matthew Spencer, and two other anonymous reviewers for providing valuable comments and suggestions that improved this manuscript. T.P. is supported by a grant from the Israel Science Foundation (878/09), by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606, and by a Recanati research grant (Tel-Aviv University). O.C. is a fellow of the Edmond J. Safra program in bioinformatics.

Literature Cited

Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 6:431–440.

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424.

Bollback J. 2005. Posterior mapping and posterior predictive distributions. In: Nielsen R, editor. Statistical methods in molecular evolution. New York: Springer. p. 439–462.

Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SG. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc Natl Acad Sci U S A. 101:9722–9727.

Carmel L, Wolf YI, Rogozin IB, Koonin EV. 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 17:1034–1044.

Charlebois RL, Doolittle WF. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. 14:2469–2477.

Ciccarelli FD, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss mapping engine. Bioinformatics 26:2914–2915.
Cohen O, Pupko T. 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. Mol Biol Evol. 27:703–713.
Cohen O, Rubinstein ND, Stern A, Gophna U, Pupko T. 2008. A likelihood framework to analyse phyletic patterns. Philos Trans R Soc Lond B Biol Sci. 363:3903–3911.
Cordero OX, Snel B, Hogeweg P. 2008. Coevolution of gene families in prokaryotes. Genome Res. 18:462–468.
Csuros M. 2006. On the estimastion of intron evolution. PLoS Comput Biol. 2:e84.
Daubin V, Lerat E, Perriere G. 2003. The source of laterally transferred genes in bacterial genomes. Genome Biol. 4:R57.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Biol. 27:401–410.
Felsenstein J. 1992. Phylogenies from restriction sites: a maximum-likelihood approach. Evolution 46:159–173.
Gal-Mor O, Finlay BB. 2006. Pathogenicity islands: a molecular toolbox for bacterial virulence. Cell Microbiol. 8:1707–1719.
Georgiades K, Merhej V, El Karkouri K, Raoult D, Pontarotti P. 2011. Gene gain and loss events in *Rickettsia* and *Orientia* species. Biol Direct. 6:6.
Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 3:679–687.
Gray RD, Atkinson QD. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426:435–439.
Graybeal A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst Biol. 43:174–193.
Hao W, Golding GB. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 16:636–643.
Hao W, Golding GB. 2008a. High rates of lateral gene transfer are not due to false diagnosis of gene absence. Gene 421:27–31.
Hao W, Golding GB. 2008b. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. BMC Genomics 9:235.
Heath TA, Zwickl DJ, Kim J, Hillis DM. 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst Biol. 57:160–166.
Holden MT, et al. 2004. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci U S A. 101:9786–9791.
Huelsenbeck JP, Nielsen R, Bollback JP. 2003. Stochastic mapping of morphological characters. Syst Biol. 52:131–158.
Jin Q, et al. 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res. 30:4432–4441.
Kettler GC, et al. 2007. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet. 3:e231.
Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol. 28:1057–1074.
Konstantinidis KT, Tiedje JM. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A. 101:3160–3165.
Koonin EV, Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36:6688–6719.
Koski LB, Morton RA, Golding GB. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol. 18:404–412.
Lawrence JG. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol. 2:519–523.
Lawrence JG, Ochman H. 1998. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A. 95:9413–9417.
Lercher MJ, Pal C. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol. 25:559–567.
Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 405:442–451.
Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol. 56:391–412.
Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. 3:2.
Moran NA. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. Curr Opin Microbiol. 6:512–518.
Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science 323:379–382.
Nei M, Tajima F. 1985. Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. Mol Biol Evol. 2:189–205.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst Biol. 51:729–739.
Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 37:1372–1375.
Pennisi E. 2004. Microbiology. Researchers trade insights about gene swapping. Science 305:334–335.
Pol D, Siddall ME. 2001. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17:266–281.
Ronquist F. 2004. Bayesian inference of character evolution. Trends Ecol Evol. 19:475–481.
Ruano-Rubio V, Poch O, Thompson JD. 2009. Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. BMC Bioinformatics 10:383.
Sankoff D. 1975. Minimal mutation trees of sequences. SIAM J Appl Math. 28:35–42.
Sicheritz-Ponten T, Andersson SG. 2001. A phylogenomic approach to microbial evolution. Nucleic Acids Res. 29:545–552.
Simmons MP, Ochoterena H. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst Biol. 49:369–381.
Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12:17–25.
Spencer M, Sangaralingam A. 2009. A phylogenetic mixture model for gene family loss in parasitic bacteria. Mol Biol Evol. 26:1901–1908.
Swofford DL, et al. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol. 50:525–539.
Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.
Templeton AR. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 37:221–244.
Wang B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. J Mol Evol. 53:244–250.

Yang Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol. 42:294–307.
Yerrapragada S, Siefert JL, Fox GE. 2009. Horizontal gene transfer in cyanobacterial signature genes. Methods Mol Biol. 532: 339–366.
Zhaxybayeva O, Nesbo CL, Doolittle WF. 2007. Systematic overestimation of gene gain through false diagnosis of gene absence. Genome Biol. 8:402.

Associate editor: David Bryant