Mahmudi et al. BMC Bioinformatics 2013, 14(Suppl 15):S10
http://www.biomedcentral.com/1471-2105/14/S15/S10

BMC Bioinformatics

PROCEEDINGS Open Access

Genome-wide probabilistic reconciliation analysis across vertebrates

Owais Mahmudi[1], Joel Sjöstrand[2], Bengt Sennblad[3], Jens Lagergren[1*]

From Eleventh Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics
Lyon, France. 17-19 October 2013

Abstract

Gene duplication is considered to be a major driving force in evolution that enables the genome of a species to acquire new functions. A reconciliation - a mapping of gene tree vertices to the edges or vertices of a species tree - explains where gene duplications have occurred on the species tree. In this study, we sample reconciliations from a posterior over reconciliations, gene trees, edge lengths and other parameters, given a species tree and gene sequences. We employ a Bayesian analysis tool, based on the probabilistic model DLRS that integrates gene duplication, gene loss and sequence evolution under a relaxed molecular clock for substitution rates, to obtain this posterior.

By applying these methods, we perform a genome-wide analysis of a nine species dataset, OPTIC, and conclude that for many gene families, the most parsimonious reconciliation (MPR) - a reconciliation that minimizes the number of duplications - is far from the correct explanation of the evolutionary history. For the given dataset, we observe that approximately 19% of the sampled reconciliations are different from MPR. This is in clear contrast with previous estimates, based on simpler models and less realistic assumptions, according to which 98% of the reconciliations can be expected to be identical to MPR. We also generate heatmaps showing where in the species trees duplications have been most frequent during the evolution of these species.

Introduction

Phylogenetics - traditionally a field primarily concerned with inferring tree-like evolution of species - has recently been superseded by *phylogenomics* - which also includes the evolution of genomes and their functional elements, in particular the genes, in relation to the species evolution. This genomics evolution is for many areas of biology, e.g., molecular biology, the final goal, to which the species evolution then is a means. In particular, evolution of genes across species is a result of evolutionary processes such as gene duplication and loss, which in eukaryotes have been shown to be major driving forces in gene evolution.

Goodman et al [1] pioneered the field by introducing the notion of a *reconciliation* of the evolutionary history of a gene family, represented by a gene tree, with that of the corresponding species, represented by a species tree. In general, a reconciliation is a mapping of the gene tree vertices onto the species tree. Each internal vertex of the gene tree is either mapped to (1) a species tree vertex, which implies that the gene tree vertex represents a speciation or (2) a species tree edge, which implies that the gene tree vertex represents a duplication. Goodman et al. used a parsimony approach and gave an algorithm that finds the most parsimonious reconciliation (MPR), i.e., the unique mapping that explains the difference between the gene and species trees using a minimum number of duplications [1].

Arvestad et al. [2] introduced the first probabilistic model for gene evolution, which explains how a gene family evolves inside a species tree by undergoing operations such as gene duplications and gene losses.

* Correspondence: jensl@csc.kth.se
[1]School of Computer Science and Communications, KTH Royal Institute of Technology, Science for Life Laboratory (SciLifeLab), Swedish e-Science Research Centre, Stockholm, Sweden
Full list of author information is available at the end of the article

Later Arvestad et al. [3] proposed an integrated model of gene duplication, gene loss, and sequence evolution, under a molecular clock for estimating the posterior distribution over gene trees. A Markov Chain Monte Carlo (MCMC) based approach was used to get the posterior distribution over gene trees, given sequence data for a gene family and the corresponding species tree. Åkerborg et al. [4] improved the model by introducing a relaxed molecular clock for sequence evolution integrated with gene duplication and gene loss. This framework efficiently computes the posterior over gene trees. Nevertheless, they do not suggest how to obtain reconciliations from the posterior distribution over gene trees. Rasmussen et al. [5] recently introduced another probabilistic approach to reconstruct gene trees inside the species tree. The method uses a hill-climbing-based approach, but it only considers MPR while computing the likelihood of a gene tree. They supported this assumption by a simulation study, where they simulated reconciled gene trees for the species tree using independently estimated duplication and loss rates [6], and found that 98% of all generated reconciliations were identical to MPR. Doyon et al. [7] reported similar results, and concluded that the most likely reconciliation is either identical to MPR or very close to MPR. In a recent study, Doyon et al. [8] using a simple birth-death process and realistic but averaged gene duplication/loss rates, found that a very small subset of all reconciliations needs to be explored in order to approximate the posterior probability of the most likely reconciliations. Åkerborg et al. [4], on the other hand argued that MPR provides an incorrect explanation of the evolutionary history of gene families that have a higher duplication rate.

Recently, genomes of different species have been published with increasingly better coverage. For instance, Heger et al. [9] published the Orthologous and Paralogous Transcripts in Clades (OPTIC) database, which provides sets of gene prediction, gene families, and orthology assignments for clades of amniotes, vertebrates, flies, nematodes and yeasts. In this study, we extend the framework by Åkerborg et al. [4], for computing the posterior over gene trees, by proposing algorithms for sampling reconciliations as well as computing the most likely reconciliations on the vertebrates clade of OPTIC dataset. This allows us to perform a genome-wide study on the OPTIC dataset, in which posteriors over gene trees *and* reconciliations are estimated. We augment the species tree by adding a heatmap for each edge, illustrating how frequently duplications occur on the edge, among the gene families. We also compare the reconciliations we obtain with the most parsimonious and conclude that MPR leads to an incorrect reconciliation in 19% of all reconciliations. Finally, we propose algorithms for sampling and computing the most likely *realizations* (a finer reconciliation, that maps vertices of the gene tree to specific time points on the species tree).

Methods

In this section, we review the DLRS model and some algorithmic results from [4]. We then continue to show how the latter can be extended so that reconciliation and so-called discretized realizations can be sampled from the posterior distribution, as well as *maximum aposteriori* (MAP) reconciliations and realizations can be computed. Finally, we describe how the difference between two reconciliations can be quantified.

The DLRS model and notation

The DLRS model, proposed by Åkerborg et. al. in [4], is based on three submodels: a duplication & loss model (DL), a substitution rate model (R), and a sequence evolution model (S) (see Figure 1). The duplication & loss submodel captures the evolution of a gene tree inside a species tree with given divergence times. For a tree T, we use $V(T)$, $E(T)$, and $L(T)$ to denote the set of vertices, edges, and leaves of a tree T, respectively. Along an edge $e \in E(S)$ of the species tree, gene duplications and losses are modeled by a linear birth-death process. The duplication & loss process has two rates that are used, in the natural way, as rates for the birth-death process. A relaxed molecular clock is assumed for the substitution rate submodel. The gene tree edges have substitution rates, which are independently and identically Γ-distributed and parameterized by a mean and a variance. The sequence evolution submodel, can be any standard sequence evolution model, e.g., JTT, which is the case of this study.

We use planted binary gene and species trees, i.e., the trees can be obtained by adding a new vertex, a so-called planted root, to a rooted binary tree and making the planted root and the root adjacent. Moreover, let $\theta = (\lambda, \mu, m, v, M)$ be the parameters of the model, where λ is the gene duplication rate, μ the gene loss rate, m the mean and v the variance of the distribution for sequence evolution rates across gene tree edges, and, finally, M the parameters of the sequence evolution model. Let T be a rooted tree and $u \in V(T)$. The subtree of T rooted at u, T_u, is the minimal subtree of T containing all descendants of u, including u. The subtree of T planted at u, denoted T^u is defined to be the subtree rooted at u, T_u, together with the edge from u to its parent.

MCMC estimation of DLRS posterior over gene trees

We now describe the MCMC based framework employed in [4], which uses the Metropolis-Hastings algorithm for inference of the posterior over gene trees

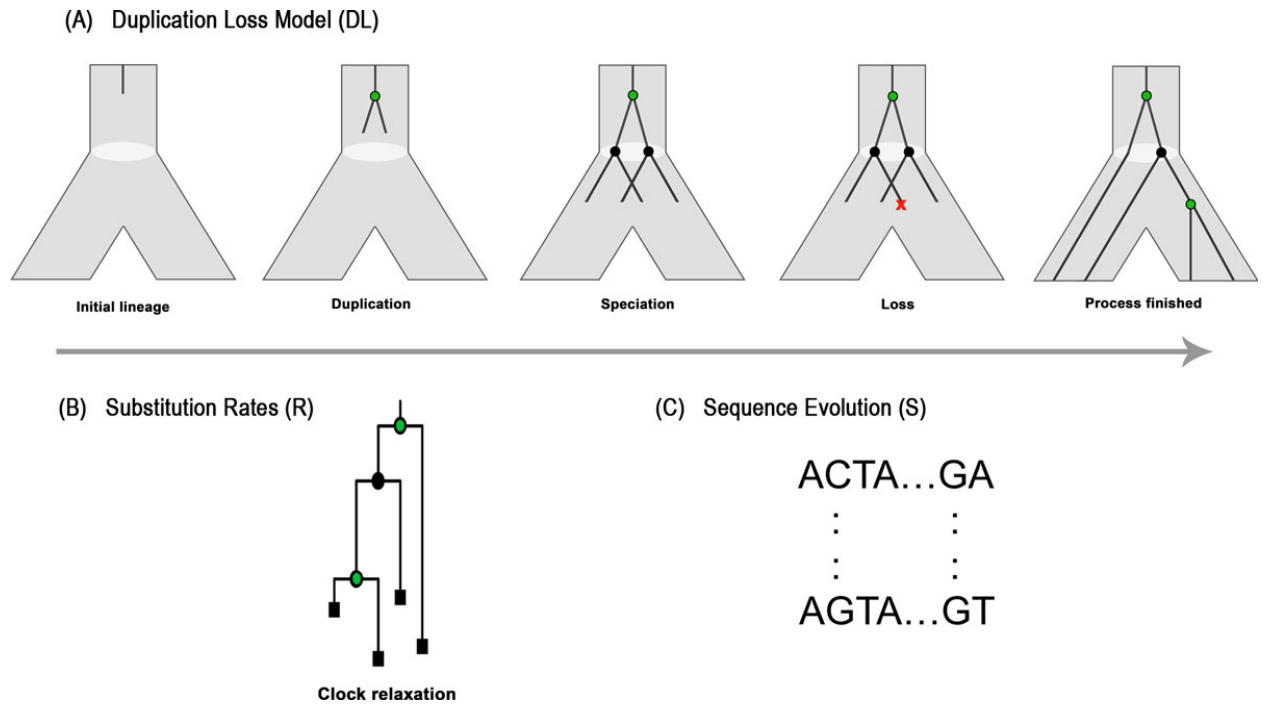

Figure 1 The three submodels of DLRS are shown. (A) Evolution of a gene lineage inside a species tree edge is modeled by a birth-death process. A gene lineage may come across a duplication event (represented by a green vertex), or a speciation event (represented by a black vertex). Every time a gene lineage passes through a speciation event, it splits into two independent gene lineages. A gene lineage may also be lost (represented by a red cross). After pruning all lost lineages, the final gene tree is obtained. (B) A relaxed molecular clock is employed to achieve branch lengths. (C) Finally, a standard sequence evolution model generates sequences over the gene tree with branch lengths.

and rate parameters, and we also show how it can be extended to facilitate sampling of reconciliations from the posterior distribution. A state of the MCMC chain is a triple (G, l, θ) where the components are: a gene tree G with lengths l, and the parameters of the DLRS model θ. We use $P(\cdot)$ to denote a probability and $p(\cdot)$ to denote a probability density. Let (G, l, θ) be the current state and let (G', l', θ') be the proposed state in an iteration of the MCMC algorithm. The acceptance probability of the proposed state (G', l', θ') is determined by the ratio of the two probability densities $p(G, l, \theta|D, S)$ and $p(G', l', \theta'|D, S)$, where D is gene sequence data and S is the species tree. Since each such can be express using Bayes equality, e.g.,

$$p(G, l, \theta|D, S) = \frac{P(D|G, l)\, p(G, l|\theta, S)\, p(\theta)}{P(D|S)},$$

the denominators cancel and we obtain

$$\frac{p(G, l, \theta|D, S)}{p(G', l', \theta'|D, S)} = \frac{P(D|G, l)\, p(G, l|\theta, S)\, p(\theta)}{P(D|G', l')\, p(G', l'|\theta', S)\, p(\theta')}.$$

Here the numerator and denominator have the same structure, so it is sufficient to describe how to compute the former. First, the factor $P(D|G, l)$ can be computed using the dynamic programming (DP) algorithm proposed by Felsenstein [10]. Second, the prior $p(\theta)$ is chosen so that it can be easily computed. Finally, the main technical contribution of [4] is a DP algorithm for computing the remaining factor $p(G, l|\theta, S)$, and we continue by outlining that approach.

Let us first define some key concepts. Let S' be a discretized species tree where edges of the species tree S have been augmented with additional discretization vertices such that all the augmented vertices are equidistant within an edge, see supplementary Figure 2 in additional file 1.

Furthermore, we define a *reconciliation* to be a mapping of vertices of a gene tree $V(G)$ to the vertices and edges of the species tree, i.e., $V(S) \cup E(S)$. A *discretized realization*, or *d-realization*, α, is a mapping of vertices of a gene tree G to the vertices of the discretized species tree S'.

A realization never maps a vertex and its parent to same vertex $x \in V(S')$. We consider sound reconciliations and realizations, e.g., they never map a vertex of the gene tree u closer to the root than the position to which it maps its parental vertex, and ensures G is properly embedded within S. Moreover, let $\sigma(u)$ be the function defined as follows (i) for a leaf $u \in L(G)$, $\sigma(u)$

is the species tree leaf in which the gene that u represents can be found and (ii) for any internal vertex u of G, $\sigma(u)$ is the least common ancestor of $L(G_u)$ in S.

Extending the posterior to d-realizations or reconciliations

In order to extend the MCMC sampling from the posterior over gene trees with lengths and parameters, i.e., $p(G, l, \theta|D, S)$ to sampling also over d-realizations, i.e., $p(G, l, \alpha, \theta|D, S)$, it is sufficient to be able to sample from $p(\alpha|G, l, \theta, S)$. This conclusion follows from

$$p(G, l, \alpha, \theta|D, S) = p(\alpha|G, l, \theta, D, S)\, p(G, l, \theta|D, S) = p(\alpha|G, l, \theta, S)\, p(G, l, \theta|D, S).$$

The analogous statement is true for a reconciliation, γ; that is, if we can sample from the reconciliation posterior distribution $p(\gamma|G, l, \theta, S)$, then we can also sample from the full posterior $p(G, l, \gamma, \theta|D, S)$.

In practice, sampling from the posterior extended with d-realizations, $p(G, l, \alpha, \theta|D, S)$, is perfomed by first running the DLRS posterior MCMC so that k samples (G_1, l_1, θ_1), ..., (G_k, l_k, θ_k) are obtained and, then for each $i \in [k]$, sample α_i from $p(\alpha_i|G_i, l_i, \theta_i, S)$. The samples from $p(G, l, \alpha, \theta|D, S)$ are, finally, $(G_1, l_1, \alpha_1, \theta_1)$, ..., $(G_k, l_k, \alpha_k, \theta_k)$.

There is a unique reconciliation associated with each realization and the posterior probability of a reconciliation is approximated by the sum of the posterior probabilities of the d-realizations associated with it. Thus, we can sample a reconciliation, from the posterior distribution over those, by sampling a d-realization, from the posterior over those, and then outputting the associated reconciliation, which easily can be computed. So by following the above described procedure and then, for each $i \in [k]$, computing the reconciliation γ_i associated with α_i, we obtain k samples $(G_1, l_1, \gamma_1, \theta_1)$, ..., $(G_k, l_k, \gamma_k, \theta_k)$ from the posterior distribution over reconciliations and other parameters.

The generation probability and d-realization sampling

In [4], a DP algorithm for computing the factor $p(G, l|\theta, S)$ was described (Figure 2). The DP makes use of a table, $s(x, y, u)$, defined as the probability that when a single gene lineage starts to evolve at the vertex $x \in V(S')$, the tree G^u is generated together with the edge lengths l and, moreover, the event corresponding to u occurs at $y \in V(S')$. Let v and w be children of u in G, and let x, y, z be vertices of $V(S')$. Let $\rho(r)$ be the probability that an edge of G has rate r. Also, let $t(x, y)$ be the time between vertices $x, y \in V(S')$. The following recursions describe how the table s can be computed:

1. If $u \in L(G)$ and $x = \sigma(u)$, $s(x, x, u) = 1$.
2. If $x \in V(S)$ and $x \neq \sigma(u)$, $s(x, x, u) = 0$.
3. If $x \in V(S) \setminus L(S)$ and $x = \sigma(u)$,

$$s(x, x, u) \left(\sum_{y \in D_L(x)} s(x, y, v) \right) \left(\sum_{y \in D_R(x)} s(x, y, w) \right),$$

where $D_L(x)$ and $D_R(x)$ are the descendants of left and right child of x in S', respectively.

4. If $x \in V(S)$ and z is a child of x such that $\sigma(L(G_u)) \subseteq L(S'_z)$ and z is an ancestor of y,

$$s(x, x, u) = p_{11}(x, z)\, \varepsilon(x, \bar{z})\, \frac{\rho\left(1\left(p(u), u\right)/t(x, y)\right)}{\rho\left(l\left(p(u), u\right)/t(z, y)\right)} s(z, y, u),$$

where $\varepsilon(x, \bar{z})$ is the probability that a gene lineage starting at x does not reach any leaf $l \in L(S'_x) \setminus L(S'_z)$. However, if $y = z$, the expression reduces to the following,

$$s(x, y, u) = p_{11}(x, y)\varepsilon(x, \bar{y})\rho(l(p(u), u)/t(x, y))s(y, y, u).$$

5. If $x \in V(S') \setminus V(S)$,

$$s(x, x, u) = 2\lambda \left(\sum_{y \in D(x) \setminus \{x\}} s(x, y, v) \right) \left(\sum_{y \in D(x) \setminus \{x\}} s(x, y, w) \right),$$

where $D(x)$ is the set of descendants of x.

6. If $x \in V(S') \setminus V(S)$ and z is the child of x in the discretized species tree S',

$$s(x, y, u) = p_{11}(x, z)\, \frac{\rho\left(l\left(p(u), u\right)/t(x, y)\right)}{\rho\left(l\left(p(u), u\right)/t(z, y)\right)} s(z, y, u).$$

However, if $y = z$, the expression reduces to the following,

$$s(x, y, u) = p_{11}(x, y)\rho(l(p(u), u)/t(x, y))s(y, y, u).$$

In any reconciliation or d-realization, the planted root of G is mapped to the planted root of S. The probability that the gene tree G is generated is the probability that when a single lineage starts at the root of S, the root of G occurs somewhere below the planted root of S and then the process continues and generates G. Hence,

$$p(G, l|\theta, S) = \sum_{y \in D(p)} s(p, y, r),$$

where p is the planted root of S, $D(p)$ its descendants, and r is the root of G. Consequently, the probability that r is mapped to $y \in V(S')$ by a d-realization sampled from all d-realizations according to the posterior probability distribution under observed G and l, is

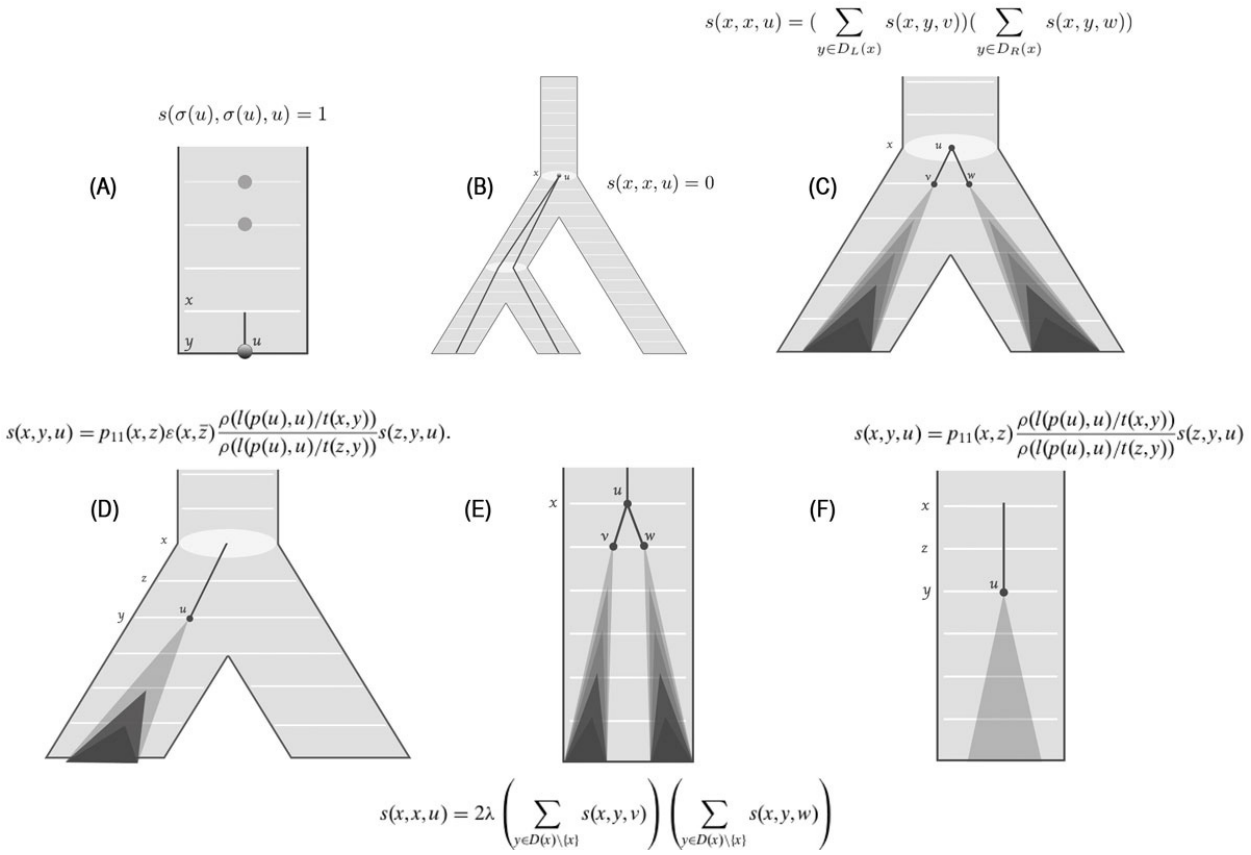

Figure 2 Recursion to compute $p(G, l|\theta, S)$. Shows six possible scenarios of a gene tree evolving inside the species tree and how they relate to dynamic programming calculations. (A) A gene tree lineage starts and reaches y, where y is a leaf of an extant species. (B) An internal vertex of a gene tree cannot be mapped to a speciation vertex other than the least common ancestor of its children in species tree. (C) An internal vertex of a gene tree may be placed on a speciation node (the least common ancestor of its children in the species tree). (D) A gene subtree rooted at u, starts from x (a speciation vertex), duplicates at y and yields the subtree below. (E) A gene subtree rooted at u is placed at some discretization point on an edge of species tree. (F) A gene lineage starts from a discretization point on an edge of species tree, and yields the gene subtree rooted at u.

$$\frac{s(p,\gamma,r)}{p(G,l|\theta,S)} = \frac{s(p,\gamma,r)}{\sum_{z\in D(p)} s(p,z,r)}.$$

Similarly, if we know that a d-realization maps a vertex $u \in V(G)$ to a vertex $x \in V(S')$, then the probability that a child v of u is mapped to y by a realization sampled from all such d-realizations, according to the posterior probability distribution under observed G and l, is

$$\frac{s(x,\gamma,u)}{\sum_{z\in D(x)} s(x,z,u)}.$$

This clearly provides an algorithm for sampling d-realizations according the posterior probability distribution under observed G and l.

Again, there is a unique reconciliation associated with each realization, and the posterior probability of a reconciliation is approximated by the sum of the posterior probabilities of the d-realizations associated with it. Thus, we can sample a reconciliation from the posterior distribution, over those, by sampling a d-realization, from the posterior over those, and then outputting the associated reconciliation.

Computing the MAP d-realization

We now give recursions that imply a DP algorithm for computing the MAP reconciliation. The following recursions describe how m can be computed, and as apparent, they are very similar to those above for s:

1. If $u \in L(G)$ and $x = \sigma(u)$, $m(x, x, u) = 1$.
2. If $x \in V(S)$ and $x \neq \sigma(u)$, $m(x, x, u) = 0$.
3. If $x \in V(S)\setminus L(S)$ and $x = \sigma(u)$,

$$m(x,x,u) = \left(\max_{y\in D_L(x)} m(x,\gamma,v)\right)\left(\max_{y\in D_R(x)} m(x,\gamma,w)\right),$$

where $D_L(x)$ and $D_R(x)$ are the descendants of left and right child of x in S', respectively.
4. If $x \in V(S)$ and z is a child of x such that $\sigma(L(G_u)) \subseteq L(S'_z)$ and z is an ancestor of y,

$$m(x,\gamma,u) = p_{11}(x,z)\,\varepsilon(x,\bar{z})\,\frac{\rho\left(l\left(p(u),u\right)/t\left(x,\gamma\right)\right)}{\rho\left(l\left(p(u),u\right)/t\left(z,\gamma\right)\right)}m(z,\gamma,u),$$

where $\varepsilon(x,\bar{z})$ is the probability that a gene lineage starting at x does not reach any leaf $l \in L(S'_x)\backslash L(S'_z)$. However, if $y = z$, the expression reduces to the following,

$$m(x,\gamma,u) = p_{11}(x,\gamma)\,\varepsilon(x,\bar{\gamma})\,\rho\left(l\left(p(u),u\right)/t\left(x,\gamma\right)\right)m(\gamma,\gamma,u).$$

5. If $x \in V(S')\backslash V(S)$,

$$m(x,x,u) = 2\lambda\left(\max_{\gamma\in D(x)\backslash\{x\}} m(x,\gamma,v)\right)\left(\max_{\gamma\in D(x)\backslash\{x\}} m(x,\gamma,w)\right),$$

where $D(x)$ is the set of descendants of x.
6. If $x \in V(S')\backslash V(S)$ and z is the child of x in the discretized species tree S',

$$m(x,\gamma,u) = p_{11}(x,z)\,\frac{\rho\left(l\left(p(u),u\right)/t\left(x,\gamma\right)\right)}{\rho\left(l\left(p(u),u\right)/t\left(z,\gamma\right)\right)}m(z,\gamma,u).$$

However, if $y = z$, the expression reduces to the following,

$$m(x,\gamma,u) = p_{11}(x,\gamma)\rho(l(p(u),u)/t(x,\gamma))m(\gamma,\gamma,u).$$

We now get an expression for the probability of the MAP d-realizations, very similar to that of $p(G, l|\theta, S)$,

$$\max_{\alpha}\frac{p(G,l,\alpha|\theta,S)}{p(G,l|\theta,S)} = \max_{\gamma\in D(p)}\frac{m(p,\gamma,r)}{p(G,l|\theta,S)}.$$

When computing the probability of the MAP d-realizations, we can use the standard technique of backpointers, i.e., keep track of the subsolution that gives the maximum value, and after the computation of m, trace the backpointers in order to find a MAP d-realization.

Posterior probability of a given reconciliation

We now give recursions for computing the posterior probability of a given reconciliation γ, i.e., $p(G, l, \gamma|\theta, S)$. The reconciliation is a mapping from $V(G)$ to $V(S) \cup E(S)$. Let R be the function from $V(S) \cup E(S)$ to $V(S')$ defined by (i) for $x \in V(S)$, $R(x) = x$ and (ii) for $(x, y) \in E(S)$, $R(x, y)$ is the set of internal vertices on the unique path between x and y in S'.

1. If $u \in L(G)$ and $x = \gamma(u)$, $s(x, x, u) = 1$.
2. If $x \in V(S)$ and $x \neq \gamma(u)$, $s(x, x, u) = 0$.
3. If $x \in V(S)\backslash L(S)$ and $x = \gamma(u)$,

$$s(x,x,u) = \left(\sum_{\gamma\in D_L(x)\cap R(\gamma(v))} s(x,\gamma,v)\right)\left(\sum_{\gamma\in D_R(x)\cap R(\gamma(w))} s(x,\gamma,w)\right),$$

where $D_L(x)$ and $D_R(x)$ are the descendants of left and right child of x in S', respectively.
4. If $x \in V(S)$ and z is a child of x such that $\sigma(L(G_u)) \subseteq L(S'_z)$ and z is an ancestor of y,

$$s(x,\gamma,u) = p_{11}(x,z)\,\varepsilon(x,\bar{z})\,\frac{\rho\left(l\left(p(u),u\right)/t\left(x,\gamma\right)\right)}{\rho\left(l\left(p(u),u\right)/t\left(z,\gamma\right)\right)}s(z,\gamma,u),$$

where $\varepsilon(x,\bar{z})$ is the probability that a gene lineage starting at x does not reach any leaf $l \in L(S'_x)\backslash L(S'_z)$. However, if $y = z$, the expression reduces to the following,

$$s(x,\gamma,u) = p_{11}(x,\gamma)\varepsilon(x,\bar{\gamma})\rho(l(p(u),u)/t(x,\gamma))s(\gamma,\gamma,u).$$

5. If $x \in V(S')\backslash V(S)$,

$$s(x,x,u) = 2\lambda\left(\sum_{\gamma\in D(x)\cap R(\gamma(v))\backslash\{x\}} s(x,\gamma,v)\right)\left(\sum_{\gamma\in D(x)(x)\cap R(\lambda(w))} s(x,\gamma,w)\right),$$

where $D(x)$ is the set of descendants of x.
6. If $x \in V(S')\backslash V(S)$ and z is the child of x in the discretized species tree S',

$$s(x,\gamma,u) = p_{11}(x,z)\,\frac{\rho\left(l\left(p(u),u\right)/t\left(x,\gamma\right)\right)}{\rho\left(l\left(p(u),u\right)/t\left(z,\gamma\right)\right)}s(z,\gamma,u).$$

However, if $y = z$, the expression reduces to the following,

$$s(x,\gamma,u) = p_{11}(x,\gamma)\rho(l(p(u),u)/t(x,\gamma))s(\gamma,\gamma,u).$$

Finally,

$$p(G,l,\gamma|\theta,S) = \sum_{\gamma\in D(p)\cap R(\gamma(r))} s(p,\gamma,r),$$

where p is the planted root of S, $D(p)$ its descendants, and r is the root of G.

Comparing reconciliations

We are interested in quantifying the difference between two reconciliations γ and γ' of G and S, in particular between a reconciliation we have sampled from the posterior and the MPR. To this end, we introduce two distance measures. First, however, an atomary distance between objects in $V(S) \cup E(S)$ is defined, so that for

any vertex $u \in V(G)$, the distance between $\gamma(u)$ and $\gamma'(u)$ is well-defined. We can then compute the two reconciliation distance measures, namely (i) the maximum atomary distance over vertices of G, and (ii) the average atomary distance over vertices of G (Figure 3).

Assume that $a, b \in V(S) \cup E(S)$. Let l be the length of the minimum length path of S that contains both a and b, and let $d(a, b) = l + 1 - |\{a, b\} \cap V(S)|/2 - |\{a, b\} \cap E(S)|$. So, for instance, if a is a vertex and b is the edge to the parent of a, then $d(a, b) = 1 + 1 - 1/2 - 1 = 0.5$, and if $a = (x, p_S(x))$ (where $p_S(\cdot)$ denotes the parent function in S) and $b = (p_S(x), p_S(p_S(x)))$, then $d(a, b) = 2 + 1 - 1 - 1 = 1$. We are now ready to define our distances between reconciliations: the max distance is

$$distance_{max}(\gamma, \gamma') = max_{u \in V(G)} d(\gamma(u), \gamma'(u)),$$

and the average distance is

$$distance_{avg}(\gamma, \gamma') = \frac{\sum_{\mu \in V(G)} d(\gamma(u), \gamma'(u))}{|V(G) \setminus L(G)|}.$$

Data

See Supplementary Material & methods (additional file 1).

Results

We applied our methods to the vertebrates clade of the OPTIC dataset, [9], consisting of the following nine vertebrate species: *Tetraodon nigroviridis* (pufferfish), *Monodelphis domestica* (gray short-tailed opossum), *Canis familiaris* (dog), *Mus musculus* (house mouse), *Homo sapiens* (human), *Ornithorhynchus anatinus* (platypus), *Taeniopygia guttata* (zebra finch), *Gallus gallus* (red junglefowl), and *Anolis carolinensis* (carolina anole). After basic filtering, 13812 gene families were selected for analysis, see supplementary Material and methods in additional file 1.

For each gene family, using the MCMC-based analysis tool PrIME-DLRS [4][11], a posterior distribution was obtained over gene trees, edge lengths and other parameters, given gene sequences and the species tree. The expected number of duplications under the posterior distribution, given the gene families and the species tree, was then estimated by sampling d-realizations and recording the number of duplications occurring at any specific discretization vertex. The number of duplications for all discretization vertices of the species tree were then normalized to 11 levels and each level was assigned a specific color. So, the colored heatmap illustrates how frequent duplications have been across the species tree. We also investigated enrichment of functional categories among the gene families with higher expected number of duplications over an edge. Finally, the appropriateness of MPR was investigated, by estimating the expected average and maximum distance, respectively, to MPR over reconciliations sampled from the posterior; a few families for which MPR was found to be unsuitable were analyzed further.

Heatmaps

Heatmaps of the number of the duplications for the posterior distribution over realizations were generated, and provide a visualization of the duplication patterns across the edges of the species tree, Figure 4A.

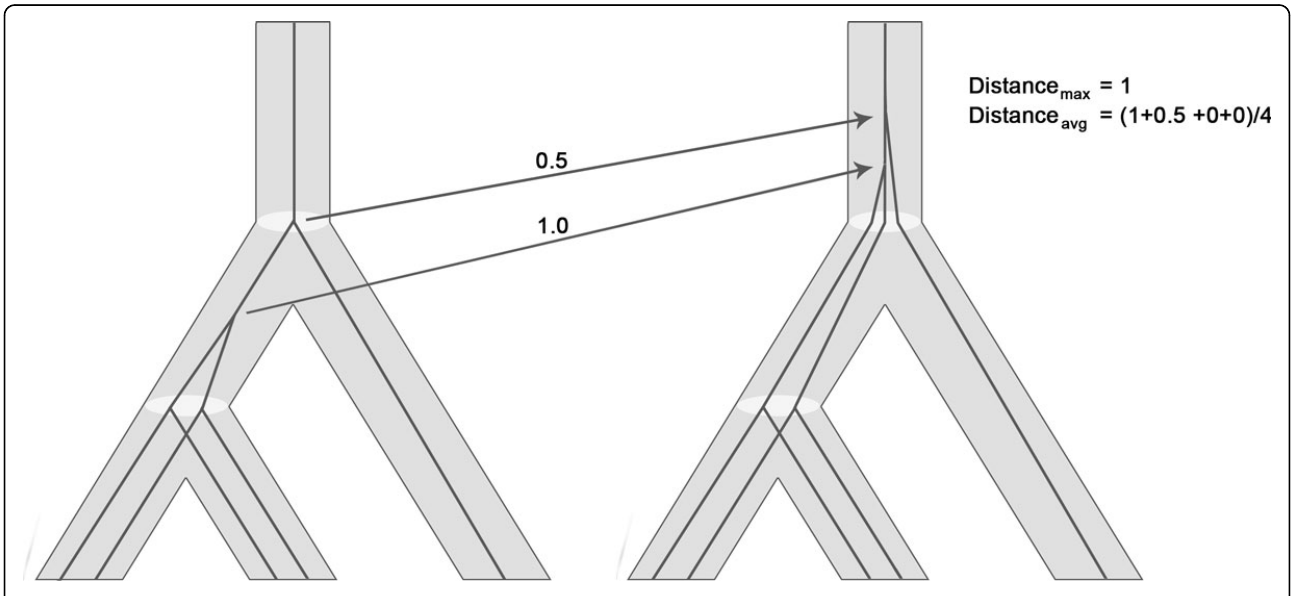

Figure 3 Computing distance between two reconciliations. Illustrates the computation of maximum distance and average distance between two reconciliations that differ in the placement of two vertices.

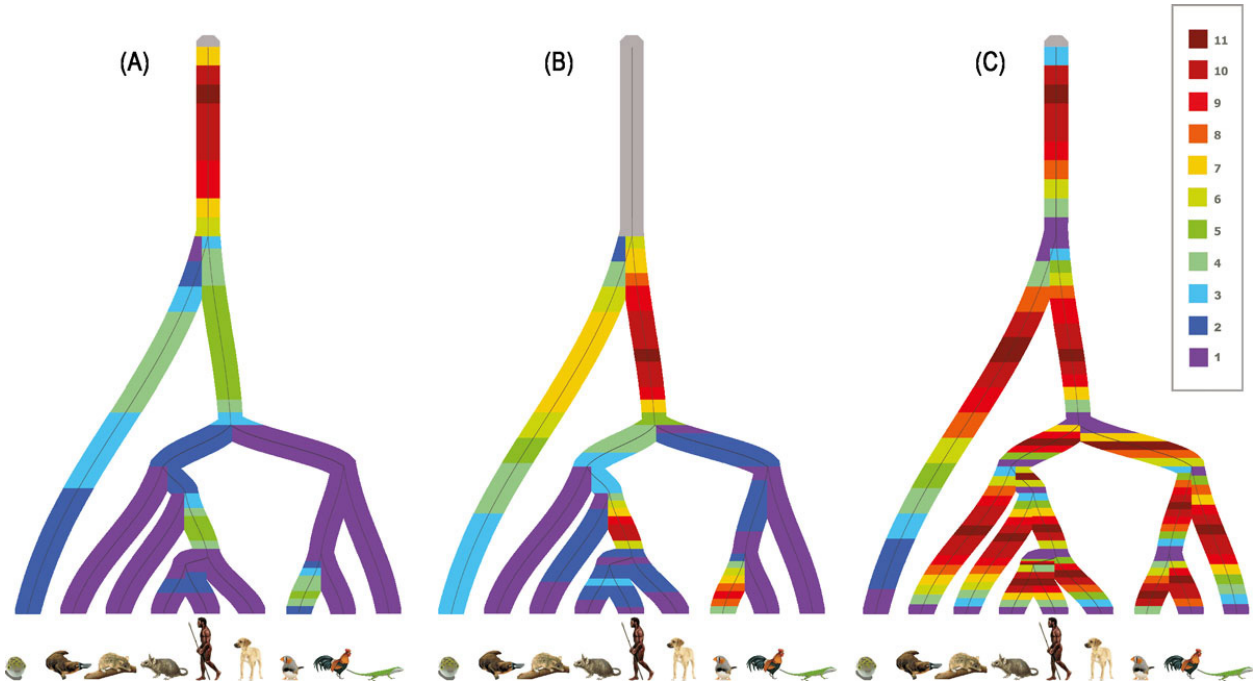

Figure 4 Heatmaps of duplications across the discretized edges of the species tree. (A) The heatmap is generated after normalizing the duplications across the tree using eleven different colors. (B) The heatmap is generated after normalizing the duplications across the tree without including the common ancestral edge. (C) The heatmap is generated by normalizing the duplications for each edge of the species tree.

The highest number of duplications were observed at the common ancestral edge of all the species. This could be interpreted as support of the 2R hypothesis proposed by Ohno [12], which suggests that the genome of early vertebrates underwent two whole genome duplications. An alternative explanation could be that incorrect gene trees in the posterior distributions give rise to duplication that reconciliations tend to place close to the planted root. In order to test the latter, we performed a *Maximum aposteriori* (MAP) analysis based on only gene families with MAP gene trees having posterior probability grather than or equal to 0.5. Heatmaps based on this data, supplementary Figure 1 (supplementary results in additional file 1), showed the same trend.

The common ancestral edge of all the species except Puffer Fish had the second highest number of duplications among all the edges of the species tree as shown in Figure 4A. In order to study the more recent lineages more closely, we normalized the duplications across the species tree without the discretization points of common ancestral edge, see Figure 4B. As the figure shows, the common ancestral edge of Human, Dog, and Mouse as well as the edge leading to Zebra Finch have comparatively higher frequencies of duplications. The higher frequencies of duplication on the edge leading to Boreoeutheria (ancestral edge of Human, Dog and Mouse) was also reported recently by Boussau et al [13].

We decided to explore the families that contribute to these duplications. Tools for performing enrichment analysis allows analysis of extant but not ancestral species and, moreover, when studying duplicating genes, choosing representative genes in extant species is complicated by the fact that the number of representatives can be varied. We, therefore, decided to work with gene families, rather than genes, and implemented this by using a single representative gene for each family. We selected the representative gene for an edge if its gene family were found to be likely to duplicate at least once on the edge. This set of genes was then annotated using the Functional Annotation Clustering (DAVID) [14,15] tool given the background of all the representative genes of the gene families of the dataset. For the common ancestral edge of Human, Dog, and Mouse, the following clusters had a Benjamini-Hochberg false discovery rate less than or equal to 0.01: *ATP Binding, Chromosome Segregation/Mitosis, ATPase activity coupled to movement of substances, Drug Metabolism, Helicase Activity, Mitotic Sister Chromatid Segregation, Fatty Acid Metabolism/Tryptophan Metabolism* and *Death-like Domain*. Using the same criteria for the ancestral edge of Zebra Finch, gave the following clusters: *Transit*

Peptide, Mitochondrion, ATP Binding, Flavoprotein and *Nucleotide Phosphatebinding region*.

In order to know how the frequency of duplications vary across each edge of the species tree, the heatmap was also normalized across edges (see Figure 4C). In most cases, there is a unimodal behavior of an individual edge. This may be explained by the relatively low number of discretization vertices and also that the signal in the sequence data may not be strong enough to reveal a more complex trend. For individual gene families, however, more complex trends are exhibited.

Distance from MPR

We computed two distances, i.e., the average distance and maximum distance from MPR, over the posterior distribution. The distribution of average distance between the sampled reconciliations and the MPR is shown in the Figure 5. The MPRs dominates the distribution of sampled reconciliations with approximately 81% of all sampled reconciliations. We computed the expected distance for posterior reconciliations of individual gene families to MPR, in order to identify gene families with a clear signal for early duplications, i.e., early in the sense of being inferred as significantly earlier than according to MPR.

The expected distance of a family was estimated as the expected average distance to MPR over reconciliations sampled from the posterior. The results showed that a number of gene families have a higher expected distance from the MPR, which means that MPR does not explain the true evolutionary history well in those cases. About 13% of all families had expected maximum distance equal to or greater than 0.5. The distribution of the expected maximum distance and the expected average distance of the gene families from the MPR are shown in Figure 6.

We selected four gene families for further analysis. They had a clear signal for early duplications and at least one gene from every species of the dataset. One of the four selected gene families was *Short chain dehydrogenase*. It has a clear signal in favor of non-MPR reconciliations as shown in Figure 7. Most of the reconciliations sampled for this family had average distances between 0.5 and 0.6, comprising around 74% of all sampled reconciliations. For this gene family, not a single reconciliation sampled was identical to the MPR. This gene family is annotated as steroid hormone biosynthesis.

Conclusion

We have presented methods for sampling and computing MAP reconciliations as well as d-realizations. Using these methods and the OPTIC dataset, we have provided the first biologically realistic estimate of the appropriateness of MPR. It was found that one can expect approximately

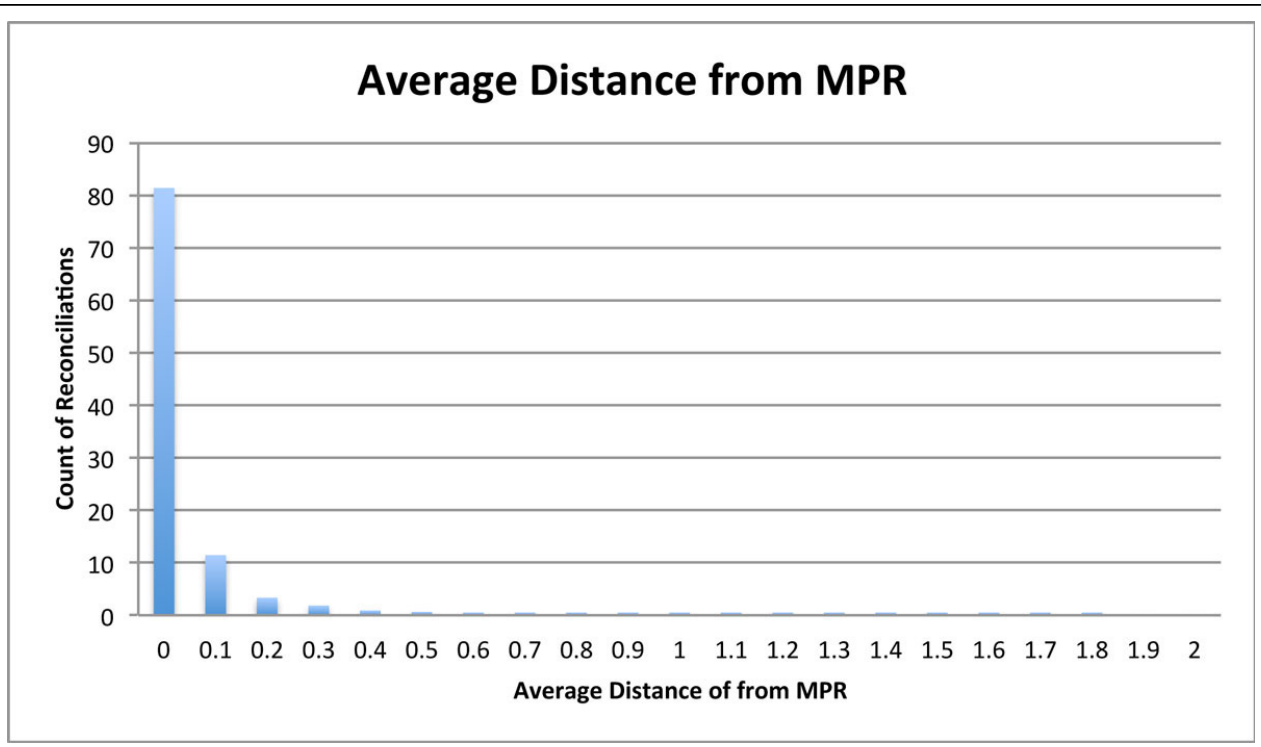

Figure 5 Average distance from MPR across gene families. X-axis represents average distance of a reconciliation from the MPR. Y-axis represents the percentage of count of reconciliations having a certain average distance from the MPR

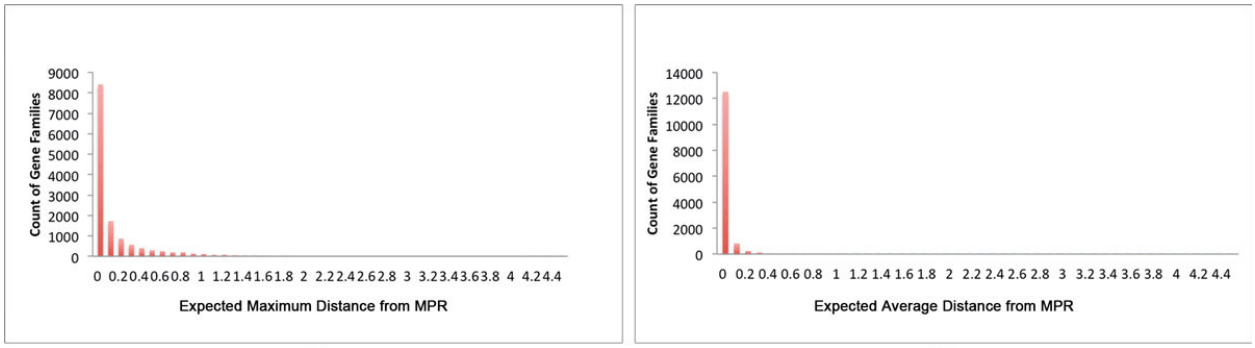

Figure 6 Expected distances from the MPR for the 13812 analyzed gene families. (A) Distribution of the expected maximum distance to MPR. (B) Distribution of the expected average distance to MPR..

Figure 7 Average distance from MPR for the gene family *Short chain dehydrogenase*. (A) Distribution of the average distance to MPR over 10000 sampled reconciliations. (B) The pie chart shows the shares of reconciliations having a certain average distance from the MPR. The labels shows the distance from the MPR.

19% of reconciliations to be different from MPR. Also, 13% of gene families can be expected to have a maximum distance of greater than or equal to 0.5 to the MPR. Among other reasons, this is interesting because some gene tree reconstuction algorithms evaluate gene trees using only MPR. We have also shown how, based on our methods, heatmaps can be constructed that illustrates how frequent duplications are across the species tree and that for vertebrates such a strategy identifies two recent edges as having hosted frequent duplications. Finally, enrichment analysis can identify functional classes among gene families that are duplicated on a specific species tree edge.

Additional material

Additional file 1: (PDF)

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study, and its publication were supported by the Swedish e-Science Research Center, The Swedish Research Council (2010-4757) and University of Engineering and Technology, Peshawar, Pakistan. BS's position is supported by a Karolinska Institute distinguished professor award to Anders Hamsten. BS acknowledge funding from the Magnus Bergvall Foundation and the Foundation for Old Servants.

This article has been published as part of BMC Bioinformatics Volume 14 Supplement 15, 2013: Proceedings from the Eleventh Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/14/S15.

Authors' details

[1]School of Computer Science and Communications, KTH Royal Institute of Technology, Science for Life Laboratory (SciLifeLab), Swedish e-Science Research Centre, Stockholm, Sweden. [2]Department of Numerical Analysis and Computer Science, Stockholm University, Science for Life Laboratory (SciLifeLab), Stockholm, Sweden. [3]Artherosclerosis research unit, Department of Medicine, Karolinska Institute, Science for Life Laboratory (SciLifeLab), Stockholm, Sweden.

Published: 15 October 2013

References

1. Goodman M, Czelusniak J, Moore GW, Romero-Herrera A, Matsuda G: Fitting the gene lineage into its species lineage, a parsimony strategy

illustrated by cladograms constructed from globin sequences. *Systematic Biology* 1979, **28(2)**:132-163.
2. Arvestad L, Berglund AC, Lagergren J, Sennblad B: **Bayesian gene/species tree reconciliation and orthology analysis using MCMC.** *Bioinformatics* 2003, **19(suppl 1)**:i7-i15.
3. Arvestad L, Berglund AC, Lagergren J, Sennblad B: **Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution.** *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology* ACM; 2004, 326-335.
4. Åkerborg Ö, Sennblad B, Arvestad L, Lagergren J: **Simultaneous Bayesian gene tree reconstruction and reconciliation analysis.** *Proceedings of the National Academy of Sciences* 2009, **106(14)**:5714-5719.
5. Rasmussen MD, Kellis M: **A Bayesian approach for fast and accurate gene tree reconstruction.** *Molecular Biology and Evolution* 2011, **28**:273-290.
6. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N: **Estimating the tempo and mode of gene family evolution from comparative genomic data.** *Genome Research* 2005, **15(8)**:1153-1160.
7. Doyon JP, Chauve C, Hamel S: **Space of gene/species trees reconciliations and parsimonious models.** *Journal of Computational Biology* 2009, **16(10)**:1399-1418.
8. Doyon JP, Hamel S, Chauve C: **An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework.** *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 2012, **9**:26-39.
9. Heger A, Ponting CP: **OPTIC: orthologous and paralogous transcripts in clades.** *Nucleic acids research* 2008, **36(suppl 1)**:D267-D270.
10. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *Journal of molecular evolution* 1981, **17(6)**:368-376.
11. Sjöstrand J, Sennblad B, Arvestad L, Lagergren J: **DLRS: gene tree evolution in light of a species tree.** *Bioinformatics* 2012, **28(22)**:2994-2995.
12. Ohno S, *et al*: *Evolution by gene duplication* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag; 1970.
13. Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V: **Genome-scale coestimation of species and gene trees.** *Genome research* 2013, **23(2)**:323-330.
14. Sherman BT, Lempicki RA, *et al*: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic acids research* 2009, **37**:1-13.
15. Da Wei Huang BTS, Lempicki RA, *et al*: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature protocols* 2008, **4**:44-57.

doi:10.1186/1471-2105-14-S15-S10
Cite this article as: Mahmudi *et al.*: Genome-wide probabilistic reconciliation analysis across vertebrates. *BMC Bioinformatics* 2013 **14**(Suppl 15):S10.