BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 24 2010, pages 3012–3019
doi:10.1093/bioinformatics/btq574

Genome analysis

Advance Access publication October 29, 2010

Bayesian sampling of genomic rearrangement scenarios via double cut and join

István Miklós[1,2,*] and Eric Tannier[3]

[1]Department of Stochastics, Rényi Institute, 1053 Budapest, Reáltanoda u. 13-15, [2]Data Mining and Search Research Group, Computer and Automation Institute, Hungarian Academy of Sciences, Budapest and [3]INRIA Rhône-Alpes; Université de Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622, Villeurbanne, France

Associate Editor: Martin Bishop

ABSTRACT

Motivation: When comparing the organization of two genomes, it is important not to draw conclusions on their modes of evolution from a single most parsimonious scenario explaining their differences. Better estimations can be obtained by sampling many different genomic rearrangement scenarios. For this problem, the Double Cut and Join (DCJ) model, while less relevant, is computationally easier than the Hannenhalli–Pevzner (HP) model. Indeed, in some special cases, the total number of DCJ sorting scenarios can be analytically calculated, and uniformly distributed random DCJ scenarios can be drawn in polynomial running time, while the complexity of counting the number of HP scenarios and sampling from the uniform distribution of their space is unknown, and conjectured to be #P-complete. Statistical methods, like Markov chain Monte Carlo (MCMC) for sampling from the uniform distribution of the most parsimonious or the Bayesian distribution of all possible HP scenarios are required.

Results: We use the computational facilities of the DCJ model to draw a sampling of HP scenarios. It is based on a parallel MCMC method that cools down DCJ scenarios to HP scenarios. We introduce two theorems underlying the theoretical mixing properties of this parallel MCMC method. The method was tested on yeast and mammalian genomic data, and allowed us to provide estimates of the different modes of evolution in diverse lineages.

Availability: The program implemented in Java 1.5 programming language is available from http://www.renyi.hu/~miklosi/DCJ2HP/.

Contact: miklosi@renyi.hu

Received on June 16, 2010; revised on September 28, 2010; accepted on October 6, 2010

1 INTRODUCTION

Although genome rearrangement was the very first clearly stated computational problem in biology introduced by Sturtevant and Novitski (1941), and many computational methods finding one most parsimonious rearrangement scenario have been published in the past 15 years, the statistical tools available for analysing genome rearrangements are still very limited. Hannenhalli and Pevzner (1999) published the first polynomial running time algorithm that finds a most parsimonious scenario transforming one genome into another using reversals. The method has been extended to transforming multichromosomal genomes by reversals and translocations (Hannenhalli and Pevzner, 1995). This general model is called the Hannenhalli–Pevzner (HP) model. Having a single parsimonious scenario is usually not enough to draw conclusions on the evolution of the genomes. For example, Bergeron *et al.* (2008) have computed different values of the breakpoint reuse rate for different scenarios transforming a human genome into a mouse genome, which tell different stories about their evolution. Exact counting or sampling in the space of parsimonious HP scenarios seems to be difficult and has only been achieved for very small examples (Braga *et al.*, 2008). Several authors published Markov chain Monte Carlo methods that sample from a Bayesian distribution of genome rearrangement scenarios (Durrett *et al.*, 2004; Larget *et al.*, 2002; Miklós, 2003), but these Markov chains have very poor mixing (Darling *et al.*, 2008), and it has been proved that there are series of data (pair of genomes) for which the mixing time of these Markov chains grows exponentially with the size of the data (Miklós *et al.*, 2010). This means that billions of Markov chain Monte Carlo (MCMC) steps are necessary to get a sufficient number of low-correlated samples from the Markov chain, making the approach computationally infeasible for large datasets. Miklós and Darling (2009) introduced a massively parallel MCMC method that samples from the uniform distribution of most parsimonious reversal scenarios, and approximately counts the number of such scenarios. By applying the method on real biological data, they showed that it is superior to other available methods. However, they could not prove that fast mixing of their Markov chain is guaranteed for any input. Furthermore, their method works only for most parsimonious reversal scenarios, and it is not clear how to extend it to sampling from the Bayesian distribution of all possible HP scenarios.

The double cut and join (DCJ) model was introduced by Yancopoulos *et al.* (2005) and in a slightly different version by Bergeron *et al.* (2006). This model allows all types of mutations of the HP model, together with mutations that are rarely observed in biological data. In particular, the formation of circular chromosomes, which is frequent in the DCJ model, is never observed in the evolution of Eukaryote genomes. A few big transpositions possibly issued from such a mechanism are hypothesized (Ross *et al.*, 2005), and some cancer karyotypes carry circular chromosomes. However, observed as well as predicted rearrangements are almost always of HP type. For example, Gordon *et al.* (2009) infer a scenario with

*To whom correspondence should be addressed.

144 rearrangements on the yeast genome that we also use in this study, all of them being HP mutations. So a current convention, supported by past observations, which might be refuted in the future, assumes that the DCJ scenarios have lower biological relevance than translocations and reversals scenarios. On the other hand, the algorithmics of DCJ is much simpler. Indeed, not only can the DCJ distance be calculated in linear time (Yancopoulos *et al.*, 2005) (as well as the HP distance), but also it is possible to quickly count the number of most parsimonious DCJ scenarios of co-tailed genomes (Braga and Stoye, 2009; Ouangraoua and Bergeron, 2009). Furthermore, it is also possible to quickly sample from the uniform distribution of DCJ scenarios of co-tailed genomes (Ouangraoua and Bergeron, 2010). The computational simplicity on its own would not be sufficient to intensively study the DCJ model; however, it opened a new perspective in developing novel methods for inferring genome rearrangements. Indeed, the DCJ model has turned to be a computationally useful concept, for example, the genome rearrangement distance under the HP model can be calculated in linear time via the DCJ distance (Bergeron *et al.*, 2009).

2 APPROACH

In this article, we explore further how the DCJ model helps developing computational tools for inferring genome rearrangements. We conjecture that counting the number of most parsimonious HP scenarios is hard (#P-complete), and hence, sampling exactly from the uniform distribution of most parsimonious scenarios or from the Bayesian distribution of all possible scenarios is also hard. However, approximate stochastic counting or sampling could be possible, and one way to do it is to use the computationally simpler DCJ model.

We introduce a parallel Markov chain Monte Carlo method sampling DCJ scenarios transforming one genome into another. Each Markov chain has its fixed temperature, and the distribution of the rearrangement scenarios depends on the temperature of the chain in the following way. We define an energy function mapping from the scenarios to real numbers. Scenarios containing many DCJs that are not in the HP model get a high energy, and the minimum energy is taken on the scenarios that have only HP-type mutations. At infinite temperature, all types of mutations of the DCJ model are equally probable. The probability of DCJ mutations that are not in the HP model decreases with the temperature, and at zero temperature, only the mutations in the HP model have non-zero probability. See Section 3 for details.

We implemented the method in the Java programming language, and tested it on yeast and mammalian data. In Section 4, we show how our method can provide quantitative estimates of the ratio of different types of rearrangements. Such numbers were previously inferred by hand (Gordon *et al.*, 2009) with a lot of uncertainties, or from a single parsimonious scenario, which, given the huge number of equivalent ones, is unsound. We correct the manual estimate of Gordon *et al.* (2009) on yeast data, and show that the rates of reversal/translocation are very different among studied taxa.

In addition to this demonstration of the applicability of our novel approach, we prove in the Appendix that

(1) Any DCJ scenario can be transformed into an HP scenario using the small changes of the Markov chains, such that the energy is monotonously decreasing. This means that the surface of the energy function we defined is smooth in the sense that all local minima are global minima.

(2) With an appropriately chosen number of parallel chains and temperatures, the probability of exchanging information between any two neighbour chains is above $1/2$. Furthermore, the highest temperature can be chosen to infinite, and the lowest temperature can be chosen such that the probability of an HP scenario in its distribution is at least $\frac{1}{2}$.

The first theorem guarantees that the MCMC cannot be trapped in a local minimum. The second theorem shows that the MCMC will not be trapped even in a global minimum, it can break out from a global minimum to visit another global minimum. Such break out is essential for fast mixing, as Miklós *et al.* (2010) proved that these global minima might be far from each other, thus causing slow mixing.

In summary, we build theoretically founded and biologically sound rearrangement scenarios. This allows to observe some modes of evolution in different taxa, as well as to provide analyses for the construction of more precise models of structural evolution.

3 METHODS

3.1 Definitions and notations

3.1.1 Genomes A *gene* a is an oriented sequence of DNA, identified by its *tail* a_t and its *head* a_h. Tails and heads are the *extremities* of the genes. An *adjacency* is an unordered pair of gene extremities. A *genome* is a set of adjacencies on a set of genes, where a gene extremity belongs to at most one adjacency. Each adjacency in a genome means that two gene extremities are consecutive on the DNA molecule. In a genome, a gene extremity g which does not belong to an adjacency is called a *telomere*.

For a genome Π on a set of genes, we define the graph G_Π: its vertex set is the set of all gene extremities, and its edge set is composed of $a_t a_h$ for every gene a, plus the adjacencies of Π.

The graph G_Π is composed of disjoint cycles and paths. Each component of G_Π is called a *chromosome* of Π. A chromosome is *linear*, if it is a path, and *circular* if it is a cycle. A genome is *linear*, if it has only linear chromosomes, and *circular*, if it has only circular chromosomes.

3.1.2 DCJ and HP operations If $\mathcal{G}$ is the set of gene extremities, let $\mathcal{G}^* = \mathcal{G} \cup \{T\}$ be the same set extended with an additional null element T. A DCJ ρ is an oriented pair $((pq,rs),(pr,qs))$, where p,q,r,s are elements of $\mathcal{G}^*$, and couples pq, rs, pr, qs, as well as couples (pq,rs) and (pr,qs) are not oriented. A DCJ is *valid* for a genome Π if pq and rs are either adjacencies of Π, of type Tx where x is a telomere of Π, or of type TT. Applying a valid DCJ on Π removes the two adjacencies pq and rs if they exist, and creates the two adjacencies pr and qs or telomeres x if pr or qs equals xT. The result of applying a valid DCJ ρ on a genome Π is denoted by Π/ρ.

A *DCJ scenario* between two genomes Π and Γ on the same set of genes is a sequence $\rho_1 \ldots \rho_k$ of valid DCJ operations (ρ_i is valid for $\Pi/\rho_1/\ldots/\rho_{i-1}$) such that $\Pi/\rho_1/\ldots/\rho_k = \Gamma$. The length of a shortest DCJ scenario between two genomes Π and Γ is the *DCJ distance*, denoted by $d_{DCJ}(\Pi,\Gamma)$. Given two genomes Π and Γ, a valid DCJ operation for Π falls into one of these three categories: decreasing the distance by 1, not changing the distance and increasing the distance by 1. These three types will be denoted by -1-DCJ, 0-DCJ and +1-DCJ, respectively.

In a scenario S, two consecutive DCJs $((pq,rs),(pr,qs))$ and $((ab,cd),(ac,bd))$ are said to *commute* if the intersection between $\{p,q,r,s\}$ and $\{a,b,c,d\}$ is empty or is the singleton $\{T\}$. It is easy to check that if two consecutive DCJ commute, then it is possible to swap their positions in the scenario, and obtain a valid scenario again.

A DCJ operation can change the structure of a genome in several ways. If it acts on one or two linear chromosomes and does not create any circular chromosome, then it is called an *HP operation*. The DCJs that are also HP operations are the following: reversals, changing the order and orientation of the genes in a segment of a chromosome, reciprocal translocations, where two chromosomes exchange an arm (possibly one of the arms is empty), fusions of two linear chromosomes, and fissions of one linear chromosome into two linear ones. The DCJs that are not HP operations are: fusions of one linear and one circular chromosome or of two circular chromosomes, fissions of one chromosome into a linear and a circular chromosome or into two circular chromosomes, circularization or linearization of chromosomes.

The number of valid HP operations for a genome Π will be denoted by $v(\Pi)$.

3.1.3 Breakpoint graphs The *breakpoint graph* of two genomes Π and Γ on the same set of genes, denoted by $BP(\Pi,\Gamma)$, is the graph in which vertex set is the set of extremities of the genes, and in which there is an edge between two vertices x and y if xy is an adjacency in either Π (these are Π-edges) or Γ (Γ-edges).

The breakpoint graph of two genomes is a set of disjoint paths and cycles. The DCJ distance is immediately readable from the breakpoint graph.

THEOREM 1. *(Bergeron et al., 2006) For two genomes Π and Γ,*

$$d_{DCJ}(\Pi,\Gamma)=n-\left(c(\Pi,\Gamma)+\frac{pe(\Pi,\Gamma)}{2}\right),$$

where n is the number of genes, $c(\Pi,\Gamma)$ is the number of cycles of the breakpoint graph and $pe(\Pi,\Gamma)$ is the number of paths with an even number of edges (possibly trivial paths with zero edges).

3.2 MCMC

We apply a parallel MCMC method, where there are several Markov chains, each converging to a prescribed distribution. Each chain has a fixed hypothetical temperature, the target distribution of the chain depends on the temperature. Unlike the usual Parallel Tempering method (Geyer, 1991), the distribution is not simply a Boltzmann distribution. In Section 3.2.1, we precisely define the target distribution of a chain at a particular temperature. The chains are coupled in a Metropolis Coupling way described in Section 3.2.2.

3.2.1 The probability distribution of a chain at a particular temperature The states of the Markov chains are the tuples $\{R,r\}$, where R is a DCJ scenario between genomes Π and Γ, and r is the rate of any DCJ mutation. To reach a distribution over this state space, we first introduce a Markov model on HP scenarios.

We model genome evolution with a continuous time Markov model, where any valid HP mutation can happen with a rate r. Since we cannot measure the evolutionary time and the rate of mutations independently, w.l.o.g., we can set the evolutionary time to unit 1, and let only the evolutionary rate vary.

The probability density that the first HP mutation happens after time t is

$$e^{-v(\Pi)rt} \tag{1}$$

The probability of a particular scenario $R=\rho_1,\rho_2,\ldots,\rho_k$ starting from genome Π given parameter r is given by the integral

$$P(R|\Pi,r)=\int_{t_k=t_{k-1}}^{1}\int_{t_{k-1}=t_{k-2}}^{1}\cdots\int_{t_1=0}^{1} e^{-v(\Pi)rt_1}r\times$$
$$\times e^{-v(\Pi/\rho_1)r(t_2-t_1)}r\times\cdots\times e^{-v(\Pi/\rho_1\rho_2\ldots\rho_{k-1})r(t_k-r_{k-1})}r\times \tag{2}$$
$$\times e^{-v(\Gamma)r(1-t_k)}\mathrm{d}t_1\mathrm{d}t_2\ldots\mathrm{d}t_k$$

Although there is no closed formula for Equation (2), it can be analytically calculated by an efficient algorithm introduced by Miklós *et al.* (2004). We can define the likelihood of parameter r using the data augmentation on the possible scenarios:

$$L(r)=P(\Gamma|r,\Pi)=\sum_{R|(\Pi,\Gamma)}P(R|r,\Pi) \tag{3}$$

where $R|(\Pi,\Gamma)$ means a summation over scenarios R transforming Π into Γ. By setting the prior distribution of r to the exponential distribution with expectation 1, we arrive to the posterior distribution of $\{R,r\}$:

$$P\big(\{R,r\}\big)\propto P(R|\Pi,r)e^{-r} \tag{4}$$

We extend the formula in Equation (2) to DCJ scenarios. Note that in this case, $P(R|\Pi,r)$ is not a probability distribution in the sense that the equation

$$\sum_{\Gamma}\sum_{R|(\Pi,\Gamma)}P(R|\Pi,r)=1 \tag{5}$$

holds only if we restrict the summation to HP scenarios. However, $P(\{R,r\})$ in Equation (4) is still a well-defined distribution when R can be any DCJ scenario. Furthermore, if we restrict this extended distribution to the HP scenarios, we get back the proper Bayesian distribution. This restriction can be done by omitting the non-HP scenarios from the MCMC samples.

For each scenario, we define the DCJ energy function, $c(R)$, as the sum of the number of circular chromosomes along the scenario. The target distribution of the Markov chain is

$$\pi_i(\{R,r\})\propto P\big(\{R,r\}\big)\times e^{-\frac{c(R)}{T_i}} \tag{6}$$

where T_i is the temperature of chain i. The restriction of this distribution to the HP scenarios is exactly the Bayesian distribution defined in Equation (4), which is our aim.

3.2.2 MCMC steps There are two types of steps in the global MCMC, composed of k parallel chains:

- chain-swapping steps
- in-chain steps.

The probability of both types of steps is 0.5. In chain-swapping steps, a random i is drawn from the uniform distribution on $[1..k-1]$. The two states of the two chains i and $i+1$ are swapped with probability

$$\min\left\{1,\frac{e^{-\frac{c(R_i)}{T_{i+1}}}\times e^{-\frac{c(R_{i+1})}{T_i}}}{e^{-\frac{c(R_i)}{T_i}}\times e^{-\frac{c(R_{i+1})}{T_{i+1}}}}\right\} \tag{7}$$

Swapping the two states with this probability guarantees the convergence of the chains to their prescribed distribution (Geyer, 1991). Equation (7) shows that all hidden normalizing constants introduced by extending the HP scenarios to DCJ scenarios and by defining distributions up to unknown normalizing constants in Equation (4 and 6) cancel out.

When an in-chain step is chosen, a random i is drawn from the uniform distribution on $[1..k]$. With probability $\frac{1}{2}$, a random window w is chosen from state of chain i, and a new scenario is proposed for this window, independently from the current scenario in the window. The new scenario is constructed by iteratively proposing a new valid DCJ for the current genome. For each valid DCJ to be chosen, the type is first determined according to Table 1. Then we choose one DCJ of this type uniformly, and apply this mutation to the genome. If the result is not the genome Γ, then the outcome will be the input of the next step. If the result is Γ, then we stop with probability 0.99, and with probability 0.01 we propose Γ as the input of the next step. Since we know the cardinality of all the three subsets of mutations, we can calculate the proposal and backproposal probabilities easily. The probability of accepting the new scenario is defined by the Metropolis–Hasting ratio (Hastings, 1970; Metropolis *et al.*, 1953). With probability $\frac{1}{2}$, mutation rate r is changed following a standard Metropolis–Hastings algorithm (Liu, 2001).

The method has been implemented in Java 1.6 and its correctness has been tested on some toy examples.

Table 1. Probabilities for the sampling strategy

There are			Sampling		
−1-DCJs	0-DCJs	+1-DCJs	−1-DCJs	0-DCJs	+1-DCJs
Yes	Yes	Yes	α	β	γ
Yes	Yes	No	α	$\beta+\gamma$	0.0
Yes	No	Yes	α	0.00	$\beta+\gamma$
Yes	No	No	1.00	0.00	0.00
No	No	Yes	0.00	0.00	1.00

There are only five different cases because it is easy to prove that if there is no −1-DCJ then there is no 0-DCJ, and in this case there must be at least one +1-DCJ. In all runs, we set $\alpha=0.95$, $\beta=0.04$, $\gamma=0.01$. See text for details.

4 RESULTS

4.1 MCMC diagnosis

Theorem 3 in Appendix B provides an upper bound on the number of chains which are necessary to guarantee an intensive exchange of states between the Markov chains. In practice, we used a significantly lower number of chains, only 10 parallel chains. Also, we used a different temperature scaling than the theoretical one in the Theorem 3. The inverse of the temperature for the k-th chain was set to

$$\frac{1}{T_k}=k^2\times s \tag{8}$$

where s was set to 0.01 for the two yeast analyses and both to 0.01 and 0.04 in several runs of mammalian analyses. In this way, the temperature of the 0 index chain was set to infinity, and the coldest chain's temperature was set to 1 or 0.25. The empirical swap probabilities for the mammalian analysis are shown on Figure 1. Although the swap probabilities between high temperature chains were relatively small, there was still a reasonable swap frequency, providing hundreds of swaps in a 200 000 step long MCMC run.

While this setup for the temperatures and number of chains provided reasonable mixing of the chains, we cannot say by any mean that these protocols are optimal. In general, the smaller the difference between neighbour chains, the higher the acceptance probability of accepting a swap between two chains; furthermore, the smaller the temperature of the coldest chain, the higher the fraction of HP scenarios in it. However, a smaller coldest temperature as well as smaller differences between the temperatures of neighbour chains call for a larger number of chains, hence increased computational time to perform one in-chain step on all chains. Furthermore, the increased number of chains also increases the hitting time of the coldest chain from the hotter chain (in terms of accepted swaps).

We opted for the quadratic function in Equation (8) because according to our experience, the acceptance probabilities of swapping two chains decrease more with the temperature gap at high temperatures. The heuristic explanation for this is that the energy function varies more at high temperatures, and thus, Δc in Lemma 5 takes high values more frequently than at low temperatures.

To show that the state swaps indeed influence all Markov chains, we calculated how many MCMC steps were needed to have a given number of chains whose state was at least once in the hottest chain during the MCMC run (Fig. 2). As expected, the number of necessary steps is smaller when the temperature difference is smaller. Similar results were obtained for the yeast analysis (data not shown).

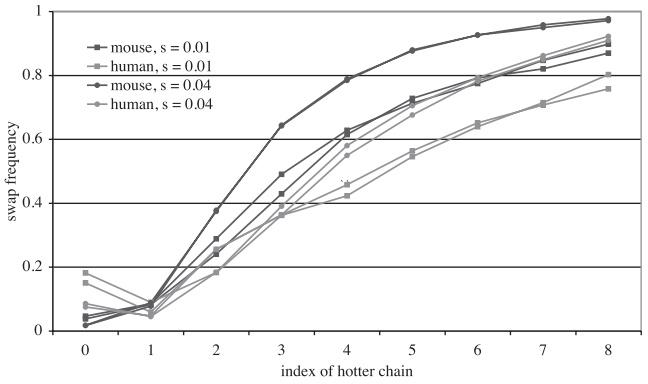

Fig. 1. The frequency of swapping the states between chains i and $i+1$ in the MCMC run. The lower indexed chain is the hotter chain.

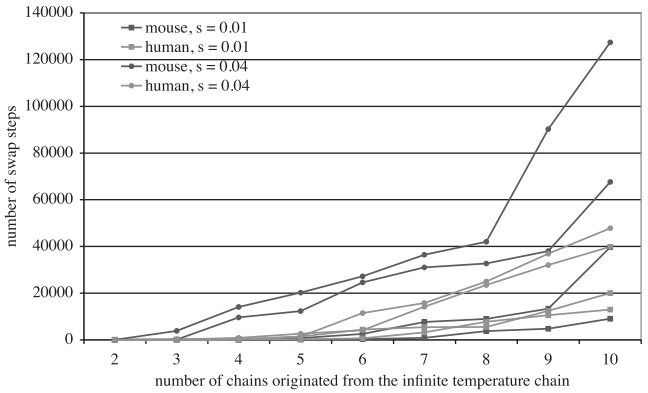

Fig. 2. The invasion of the hottest chain. See text for more details.

Each MCMC run took 200 000 MCMC steps, each chain was sampled after each 200 steps, making 1000 samples for each chain. Among these samples, we found thousands of HP scenarios ranging from 1957 till 3354 in the mammalian analyses, and above 3000, ranging up to 4736 samples in the yeast analyses, showing that the higher indexed chains were cold enough to have HP scenarios in their target distributions with non-negligible probabilities. The minimum were taken on one of the runs with $s=0.01$, showing that the HP scenarios are less frequent at high temperature.

We also would like to mention that the hottest chains almost never sampled HP scenarios, the empirical frequency of HP scenarios in the hottest chain is <0.0003. This shows that it is computationally inefficient to sample DCJ scenarios and restrict the samples to HP scenarios. Tempering is needed to sample HP scenarios frequently.

MCMC runs on each input data were repeated at least three times with different random seeds, and the convergence was also checked by comparing the samples from different runs (see, e.g. Fig. 3). showing the empirical distribution of parameter r estimated from two different runs.

4.2 Biological data analysis

We tested the MCMC sampler on two datasets. One is the comparison of the yeast *Saccharomyces cerevisiae* with its ancestral genome pre-dating the whole-genome duplication that occurred in that lineage 100 million years ago. This dataset was manually

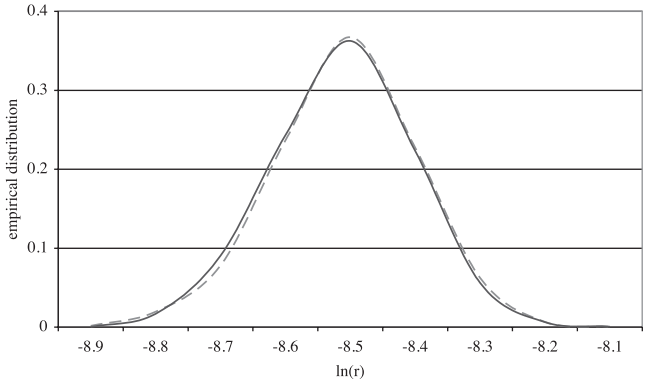

Fig. 3. The empirical distribution of the parameter r estimated from two different MCMC runs (dashed line: $s=0.01$, straight line: $k=0.01$) on the euarchontoglire-human data.

analysed by Gordon *et al.* (2009) using parsimony principles and it is interesting to compare their results with the automatic method. The other one is a human mouse genome comparison, which is interesting for two purposes. The first is a possibility of comparing the modes of structural evolution in different lineages, i.e. in yeast versus mammals. The second arises from a debate on the computation of the breakpoint reuse rate, which is sometimes computed independently from any scenario (Pevzner and Tesler, 2003) or from an *ad hoc* parsimonious scenario (Bergeron *et al.*, 2008), and has very different values according to the method, which drives opposite biological conclusions. Sampling among scenarios allows the computation of a mean breakpoint reuse.

For the yeast genome, we used as 'genes' the markers constructed by Gavranovic and Tannier (2010) using a double synteny principle. These markers cover 97% of the genes used by Gordon *et al.* (2009), from which we take the ancestral pre-duplication configuration [the ancestral genomes vary only by one DCJ (a non reciprocal translocation) from one method to the other]. Gordon *et al.* (2009) estimated the number of rearrangements from the ancestor to *S.cerevisiae* to be 73 inversions, 66 reciprocal translocations and 5 non-reciprocal translocations. Our results are summarized in Table 2. The ratio of reversals/translocations is lower than in Gordon *et al.* (2009) (0.66 versus 1.03), confirming the propensity of yeasts to evolve mainly by translocations. This is also confirmed by the comparison of the same ancestral configuration to the non-duplicated genome *Lachancea kluyveri*, where the ratio is about 24% [the manual study of Gordon *et al.* (2009) was limited to the scenario in the *S.cerevisiae* branch].

In contrast, for mammalian genomes, we used the Pecan alignment coordinates from the release 58 of the Ensembl-Compara database (Paten *et al.*, 2008), retrieving the seeds that were present in the Human, Macaca, Mouse, Rat, Dog, Horse, Cow and Opossum genomes. By joining consecutive alignments and discarding the groups covering <100 kb in at least one species, we obtained 915 universal orthology blocks covering ∼70% of the chosen mammalian genomes. We used the euarchontoglire (human–mouse ancestor) ancestral genome reconstruction using the method of Chauve and Tannier (2008), and compared its organization with the extant genomes. The results, reported on Table 2 for human and mouse, show that the modes of evolution are very different from yeast ones, and even can be very different among lineages.

Table 2. Mean numbers of rearrangements according to their types: fusion, fission, reciprocal translocation, non-reciprocal (telomeric) translocation and reversals.

	Fus.	Fis.	Rec. Transl.	Transl.	Rev.
Yeast anc. → Scerevisiae	0	0	86.5	6.7	61.3
Yeast anc. → Skluyveri	0	0	66.9	5.2	17.6
Euarch anc. → Human	3.3	0.3	7.8	10.9	61.5
Euarch anc. → Mouse	6.6	0.6	112.9	53.4	34.5

The presence of fusions and fissions contrasts with the evolution of yeast species, even if the ratio between the two is biased because of a reconstruction of an ancestral genome with 27 contiguous ancestral regions, which are probably fused into 23 or 24 proto-chromosomes in reality. The reversal/translocation ratio is surprisingly variable, the human lineage evolving mainly by reversals, whereas in mouse, as in yeasts, reciprocal translocations are the dominant mechanism.

If we count one break for a fission, one for a translocation and two for reversals and reciprocal translocations, we arrive at a breakpoint reuse rate of 1.46 for the human branch, and 1.59 for the mouse branch. Such a rate, which can have consequences on the construction of models of structural evolution, cannot be analysed on its own, its significance depends on the sizes of the chromosomal regions affected by the rearrangement breakages, thus on the properties of the constructed orthology blocks. Such an analysis is beyond the scope of this article, which nevertheless provides a way to have a good estimation of this value.

5 DISCUSSION

The estimation of rates of different rearrangements, as well as the breakpoint reuse rate, are important to devise a model of genome structural evolution, just like the estimation of a substitution matrix is necessary to use a model of nucleotide or amino acid sequence evolution. Up to now, most models are limited to one kind of rearrangements, or provide only parsimonious solutions, giving an equal probability to any kind of event. The analysis of rearrangements for distant species is thus rarely possible.

We arrived at estimations of the ratio of different kinds of rearrangements, showing that fusions and fissions are usual modes of evolution in mammalian genomes, but not in yeasts. Reversals are also much more frequent in mammals, compared with translocations. However, non-reciprocal translocations, called telomeric translocations in Gordon *et al.* (2009), are quite frequent in both taxa. The significance of the breakpoint reuse rate for refuting or not the Random Breakage Model is beyond the scope of this study, which can however propose a solution to the problem of its computation. These observations of different rates for different types of rearrangements, or the possible preferential usage of some breakpoints might be included as parameters in a more general model in a future work. We could also use this kind of work to test the hypothesis that non-HP DCJ, with or without immediate reincorporation of circular chromosomes, are indeed rare or irrelevant, by sampling from low but non-zero temperature, and compute the probabilities of the inferred scenarios. We leave this large debate to a future work.

We introduced a promising approach for inferring genome rearrangement evolution. The previous best method (Miklós and Darling, 2009) was able to sample only from the uniform distribution of all most parsimonious reversal scenarios. Our new method provides a radically different approach that can sample also from the Bayesian distribution of all possible reversal scenarios. The MCMC we constructed has a theoretical ground; however, Theorems 2 and 3 work for more particular types of data than the MCMC itself, as they are limited to unichromosomal co-tailed hurdle-free genomes, and parsimonious scenarios. However, we conjecture they are true for a more general case, if not the most general one in which the MCMC stands. This new method also has the potential to be extended to multiple genome rearrangements. So far, no efficient method is available for the Bayesian inferring of genome rearrangement phylogenies. BADGER, the best method available for Bayesian rearrangement phylogenies (Larget *et al.*, 2005) needs tens of millions of MCMC steps to converge on moderate data containing less than 80 syntheny blocks (Darling *et al.*, 2008). We conjecture that the torpid mixing of BADGER is caused by the bottlenecks in the state space discovered by Miklós *et al.* (2010). As Miklós *et al.* (2010) already realized, the usual parallel tempering that heats up the distribution towards suboptimal scenarios cannot eliminate these bottlenecks. The new heating protocol introduced in this article, which heats the distribution towards DCJ scenarios can eliminate these bottlenecks.

ACKNOWLEDGEMENTS

The authors thank Wei Xu, Krister Swenson, Haris Gavranovic, Renaud Lenne and Jens Lagergren for fruitful discussions. The anonymous referees are thanked for their constructive criticisms.

Funding: Agence Nationale pour la Recherche (ANR-08-GENM-036-01 and ANR-08-EMER-011-03 to E.T.); ECO-NET project funded by the French Ministry of Foreign Affairs. Hungarian Scientific Research Fund (NK 78439).

Conflict of Interest: none declared.

REFERENCES

Bergeron,A. *et al.* (2006) A unifying view of genome rearrangements. *Lect. Notes Comput. Sci.*, **4175**, 163–173

Bergeron,A. *et al.* (2008) On computing the breakpoint reuse rate in rearrangement scenarios. *Lect. Notes Comput. Sci.*, **5267**, 226-240

Bergeron,A. *et al.* (2009) A new linear time algorithm to compute the genomic distance via the double cut and join distance. *Theor. Comput. Sci.*, **410**, 5300–5316.

Braga,M.D.V. *et al.* (2008) Exploring the solution space of sorting by reversals with experiments and an application to evolution. *IEEE-ACM Trans. Comput. Biol. Bioinform.*, **5**, 348–356

Braga,M.D.V. and Stoye,J. (2009) Counting all DCJ sorting scenarios. *Lect. Notes Comput. Sci.*, **5817**, 36–47.

Chauve,C. and Tannier,E. (2008) A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.*, **4**, 1097–1112.

Darling,A. *et al.* (2008) Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.*, **4**, e1000128.

Durrett,R. *et al.* (2004) Bayesian estimation of genomic distance. *Genetics*, **166**, 621–629.

Gavranovic,H. and Tannier,E. (2010) Guided genome halving: provably optimal solutions provide good insights into the preduplication ancestral genome of Saccharomyces cerevisiae *Proc. PSB*, **15**, 21–30.

Geyer,C.J. (1991) Markov Chain Monte Carlo Maximum Likelihood. In Keramidas,E.M. and Selma,S.M. (eds) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation of North America Inc., Fairfax Station, VA, pp. 156–163.

Gordon,J.L. *et al.* (2009) Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. *PLoS Genet.*, **5**, e1000485.

Hannenhalli,S. and Pevzner,P.A. (1995) Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of 36th Annual Symposium on Foundations of Computer Science (FOCS'95)*, IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 581–592.

Hannenhalli,S. and Pevzner,P.A. (1999) Transforming cabbage into Turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, **46**, 1–27.

Hastings,W.K. (1950) Monte Carlo sampling methods using Markov chains and their applications. *Biometrica*, **67**, 97–109.

Larget,B. *et al.* (2002) Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *J. R. Stat. Soc. B.*, **64**, 681–695.

Larget,B. *et al.* (2005) A Bayesian analysis of metazoan mitochondrial genome arrangements. *Mol. Biol. Evol.*, **22**, 485–495.

Liu,J.S. (2001) *Monte Carlo Strategies in Scientific Computing*. Cambridge University Press, New York.

Metropolis,N. *et al.* (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1091.

Miklós,I. (2003) MCMC genome rearrangement. *Bioinformatics*, **19**, ii130–ii137.

Miklós,I. and Darling,A. (2009) Efficient sampling of parsimonious inversion histories with application to genome rearrangement in Yersinia. *Genome Biol. Evol.*, **1**, 153–164.

Miklós,I. *et al.* (2004) A 'long indel' model for evolutionary sequence alignment. *Mol. Biol. Evol.*, **21**, 529–540.

Miklós,I. *et al.* (2010) The metropolized partial importance sampling MCMC mixes slowly on minimum reversal rearrangement paths. *ACM/IEEE Trans. Comput. Biol. Bioinformatics*, to appear. Available at: http://www.computer.org/portal/web/csdl/doi/10.1109/TCBB.2009.26.

Ouangraoua,A. and Bergeron,A. (2009) Parking functions, labeled trees and DCJ sorting scenarios. *Lect. Notes Comput. Sci.*, **5817**, 24–35.

Ouangraoua,A. and Bergeron,A. (2010) Combinatorial structure of genome rearrangements scenarios. *J. Comput. Biol.*, **17**, 1129–1144.

Paten,B. *et al.* (2008) Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.,* **18**, 1814–1828.

Pevzner,P. and Tesler,G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl Acad. Sci. USA*, **100**, 7672–7677.

Ross,M.T. *et al.* (2005) The DNA sequence of the human X chromosome. *Nature*, **434**, 325–337.

Sturtevant,A.H. and Novitski,E. (1941) The homologies of chromosome elements in the genus Drosophila. *Genetics*, **26**, 517–541.

Yancopoulos,S. *et al.* (2005) Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, **21**, 3340–3346.

APPENDIX A

In this appendix, we prove in a particular case that all local minima on the energy surface defined by the sum of the number of circular chromosomes along a DCJ path are global minima. We conjecture the same result for the general case (multichromosomal genomes).

A.1. Unichromosomal hurdle-free co-tailed genomes

A *unichromosomal* genome is a genome with only one chromosome. A couple of genomes is said to be *co-tailed* if they share the same telomeres. The DCJ distance between two unichromosomal linear co-tailed genomes is $n-(c(\Pi,\Gamma)+1)$.

On unichromosomal linear genomes, DCJs are either reversals, or fusions or fissions of chromosomes involving at least one circular chromosome. A fission of a chromosome into two, immediately followed by the fusion of these two chromosomes (with different points of fusions) is called a *block-interchange*.

A Π-edge (respectively Γ-edge) of the breakpoint graph of two genomes Π and Γ is said to be *oriented* if it joins two gene heads or two gene tails. It is *unoriented* if it joins a head and a tail. A cycle is said to be oriented if it has an oriented edge.

Two cycles A,B of a breakpoint graph of unichromosomal co-tailed genomes Π and Γ are said to *cross* if the minimal path of G_Π containing the extremities of A also contains extremities from B, and vice-versa.

The *interleaving graph* of two genomes Π and Γ is the graph whose vertex set is the set of cycles of the breakpoint graph, and two cycles are linked by an edge if they cross. A *component* of Π and Γ is the set of genes extremities of the cycles of a connected component of the interleaving graph of Π and Γ if this set has cardinality at least 4.

A component is said to be *oriented* if at least one edge of a cycle of the component has an oriented edge. It is *unoriented* otherwise. Π and Γ are *hurdle-free* if they have no unoriented component.

The following two results are corollaries of the Hannenhalli–Pevzner theorem (Hannenhalli and Pevzner, 1999).

RESULT 1. *For two hurdle-free co-tailed unichromosomal genomes Π and Γ, there is a DCJ scenario of size $d_{DCJ}(\Pi,\Gamma)$ which contains only reversals.*

RESULT 2. *In any DCJ scenario between two co-tailed unichromosomal genomes Π and Γ of size $d_{\mathrm{DCJ}}(\Pi,\Gamma)$, there is no reversal involving gene extremities of an unoriented component.*

The number of different DCJ scenarios between two hurdle-free co-tailed unichromosomal genomes Π and Γ can be easily computed, while computing the number of DCJ scenarios containing only reversals, as well as only reversals and block-interchanges, are open problems.

A.2. The theorem

For a DCJ scenario S between genomes Π and Γ, denoted by $S(i)$ the DCJ at the i-th position on S. Let $\Pi_i^S = \Pi/S(1)/\ldots/S(i)$ for $0 \leq i \leq d_{\mathrm{DCJ}}$ ($\Pi_0^S = \Pi$ and $\Pi_{d_{\mathrm{DCJ}}}^S = \Gamma$). For two scenarios S_1 and S_2, we define $d(S_1,S_2)$ as the smallest integer d such that there exists k verifying:

$$\text{For all } i \notin [k,k+d-1],\quad S_1(i) = S_2(i).$$

In other words, $d(S_1,S_2) \leq d$ if it is possible to replace d consecutive DCJs of S_1, by d other DCJs to obtain S_2. For a genome Π, let $circ(\Pi)$ be the number of circular chromosomes of Π. Let $c(S) = \sum_{i=0}^{k} circ(\Pi_i^S)$ be the *score* of the scenario S. For any DCJ scenario S between two unichromosomal co-tailed genomes, $c(S) \geq 0$ and S is a scenario of reversals if and only if $c(S) = 0$.

THEOREM 2. *For two hurdle-free co-tailed unichromosomal genomes Π and Γ, let S_1 be a DCJ scenario transforming Π into Γ. There exists a finite sequence $S_1 S_2 \ldots S_k$ of DCJ scenarios, such that*

- for all i, $d(S_i,S_{i+1}) \leq 3$ and $c(S_i) \geq c(S_{i+1})$;
- S_k is a reversal scenario.

A.3. Proof of Theorem 2

Let $S_1 = \rho_1 \ldots \rho_k$ be a scenario between unichromosomal co-tailed hurdle-free genomes Π and Γ such that $c(S_1) > 0$. We prove that there always exists a finite sequence $S_1 S_2 \ldots S_l$ such that $c(S_l) < c(S_1)$ and for all i, $d(S_i,S_{i+1}) \leq 3$ and $c(S_i) \geq c(S_{i+1})$. This proves the theorem.

A DCJ scenario that is not a reversal scenario contains fissions and fusions. If in a DCJ scenario every fission is immediately followed by a fusion, then we say it is a reversal/block-interchange scenario. The first step of our proof shows that any DCJ scenario can be transformed into a scenario of this type. The second step of the proof shows that this can further be transformed into a reversal scenario.

Case 1: not all fission is immediately followed by a fusion.

Let ρ_p be the first (with minimum p) such fission in S_1. So $\Pi_{p-1}^{S_1}$ is a unichromosomal genome. The DCJ ρ_p fissions the unique chromosome of $\Pi_{p-1}^{S_1}$ into two chromosomes C_1 and C_2. Let G_1 (respectively G_2) be the set of gene extremities in C_1 (respectively C_2). Now let ρ_q be the first DCJ after ρ_p in S_1 which involves gene extremities both from G_1 and G_2. It exists as Γ is unichromosomal, and by hypothesis $q > p+1$.

By hypothesis on ρ_q, ρ_{q-1} involves either only gene extremities from G_1 or only gene extremities from G_2. Suppose w.l.o.g. that it is G_1. If ρ_q and ρ_{q-1} commute, let $\rho'_{q-1} = \rho_q$ and $\rho'_q = \rho_{q-1}$. If they do not commute, denote ρ_{q-1} by $((ab,cd),(ac,bd))$ and ρ_q by $((ac,ef),(ae,cf))$, with a,b,c,d being gene extremities of G_1 and ef being gene extremities from G_2. Now let $\rho'_{q-1} = ((ab,ef),(ae,bf))$ and $\rho'_q = ((cd,bf),(bd,cf))$. In both cases, the scenario

$$S_2 = \rho_1 \ldots \rho_{q-2} \rho'_{q-1} \rho'_q \rho_{q+1} \ldots \rho_n$$

is composed of valid DCJ operations, we trivially have $d(S_1,S_2) \leq 2$ and we also have $c(S_2) \leq c(S_1)$ because

- For all $i \neq q-1$, $\Pi_i^{S_1} = \Pi_i^{S_2}$;
- Clearly , $|circ(\Pi) - circ(\Pi/\rho)| \leq 1$ for any genome Π and DCJ ρ, so $circ(\Pi_{q-1}^{S_1}) \geq circ(\Pi_{q-2}^{S_1}) - 1$;
- The DCJ ρ'_{q-1} is a chromosome fusion so $circ(\Pi_{q-1}^{S_2}) = circ(\Pi_{q-2}^{S_2}) - 1$, yielding $circ(\Pi_{q-1}^{S_1}) \geq circ(\Pi_{q-1}^{S_2})$.

Now in S_2, ρ'_{q-1} is the first DCJ after ρ_p which involves gene extremities both from G_1 and G_2. Applying this transformation again to S_2 decreases the index q of the first DCJ after ρ_p which involves gene extremities both from G_1 and G_2. We may apply the same transformation until $q = p+1$, which means that $\rho_p \rho_{p+1}$ is a block interchange.

Applying the same transformation to every ρ_p fission which is the first DCJ of a non-block-interchange type eventually gives a reversal/block-interchange scenario.

Case 2: the scenario S_1 is a reversal/block-interchange scenario.

Note that in that case the number of circular chromosomes in a scenario — its score—is also the number of block interchanges. So the goal will be to get progressively rid of all block interchanges and arrive at a reversal scenario.

First, if for a block interchange $\rho_p \rho_{p+1}$ (both operations are DCJs), the genomes Π_{p-1}^S and Π_{p+1}^S have an oriented component, then $\rho_p \rho_{p+1}$ can easily be replaced by two reversals, as a direct consequence of Result 1. Doing this yields a scenario which is distant from the initial one of at most two, and the number of circular chromosomes is deceased by one, proving the theorem.

So we may assume that all block interchanges $\rho_p \rho_{p+1}$ are *unoriented*, which means the breakpoint graph of Π_{p-1}^S and Π_{p+1}^S has only unoriented components.

Let now ρ_r and ρ_b be a reversal and a block interchange in the scenario S. We note Π^- the genome before the application of the first event among ρ_r and ρ_b, and Π^+ the genome after the application of the last one. We say that ρ_r and ρ_b *cross* if the minimal path on G_{Π^-} between the gene extremities used by ρ_r contains extremities used by ρ_b, vice-versa, exchanging ρ_r and ρ_b. It is easy to see that if they are consecutive and do not cross, then they commute and swapping their position yields a scenario which is distant of 3 from the original one, and has the same score (the reversal stays a reversal, the block interchange stays a block interchange).

Choose now ρ_r and ρ_b which cross, and such that there are as few DCJs as possible between them in S. These exist because else the reversals and block interchange would act on different components, which is not possible because there is no unoriented component.

From now we assume that ρ_r is before ρ_b, but a symmetric reasoning yields the proof for the opposite case.

Suppose there are rearrangements between ρ_r and ρ_b. If among those rearrangements, there is a reversal which is immediately before a block interchange, then by hypothesis on ρ_r and ρ_b, they do not cross. So it is possible to swap them without changing the score of the scenario, nor the crossing properties of any pair or reversal and block interchange. Iteratively, applying this allows to assume that all reversals occur after all block interchanges between ρ_r and ρ_b.

Now ρ_r occurs before a block interchange, and if it is not ρ_b, then they do not cross. So it is possible to swap them. This has the effect of applying a block interchange to Π^-. As this block interchange does not cross any reversal applied before ρ_b, it cannot change the crossing properties of reversals and block interchanges, so it is possible to repeatedly apply this procedure while there are block interchanges between ρ_r and ρ_b, there are only reversals left. In the same way, as ρ_b occurs after a reversal, if it is not ρ_r then they do not cross and it is possible to swap them without changing the crossing properties of reversals and block interchanges. After repeating this procedure there are no rearrangement anymore between ρ_r and ρ_b and they are immediately consecutive. The following lemma yields the theorem in that case.

LEMMA 1. *In a scenario S, let ρ_r and ρ_b be, respectively, a reversal and a block interchange that are consecutive (in any order) and crossing. Then it is possible to replace them by three reversals in S.*

PROOF. By Result 2 the component of Π^- and Π^+ containing the gene extremities involved in ρ_r and ρ_b is oriented since there is a scenario with a reversal transforming Π^- into Π^+. So by Result 1, there are three reversals transforming Π^- into Π^+, which proves the result. ■

APPENDIX B

In this appendix, we prove the following theorem, showing that with a non-negligible probability, the MCMC can diversify efficiently in the solution space of HP scenarios:

THEOREM 3. *For any pair of hurdle-free, co-tailed, linear genomes Π and Γ with n genes, $k=O(n^3\log(n))$ parallel chains sampling from most parsimonious DCJ scenarios following target distributions given by*

$$\pi_i(R(\Pi,\Gamma)) \propto e^{-\frac{c(R(\Pi,\Gamma))}{T_i}} \tag{9}$$

can be defined with the following properties:

- The temperature of the first chain is infinite,
- The swapping probability between any two consecutive chains given by Equation (7) is at least $\frac{1}{2}$,
- The probability of a HP scenario in the target distribution of the k-th chain is at least $\frac{1}{2}$.

The first property provides that all most parsimonious DCJ scenarios are equally probable in the target distribution of the first chain. We proved (Miklós,I. and Tannier,E., manuscript in preparation) that it is easy to sample from this distribution with Markov chains, and if the genomes are co-tailed, exact sampling is also possible (Ouangraoua and Bergeron, 2010). The second property provides that the information change between the parallel chains is not negligible. The third property provides that a few samples from the k-th chain is sufficient to get HP scenarios. Although these together do not prove fast mixing of our method, it is definitely takes us closer to a final proof.

The theorem is proved using the following lemmas.

LEMMA 2. *For any most parsimonious DCJ scenario S between Π and Γ,*

$$c(S) \leq \frac{n(n-2)}{4} \tag{10}$$

PROOF. Since Π and Γ are linear, $circ(\Pi)=circ(\Gamma)=0$. $|circ(\Pi')-circ(\Pi/\rho)| \leq 1$ for any genome Π', thus $c(S)$ is maximal, if the number of circles increases till the middle of the path, and then decreases. Since the maximum length of the path is $n-1$, Equation (10) immediately holds. ■

LEMMA 3. *The number of most parsimonious DCJ scenarios between Π and Γ is at most $(4n^2-n)^{n-1}$.*

PROOF. n genes have $2n$ extremities, forming at most $2n$ telomeres and adjacencies. There are at most two DCJs acting on a given pair of telomeres and/or adjacencies, having an upper bound of $2\binom{2n}{2}$ DCJs acting on a pair of telomers/adjacencies. Above these, there are fissions involving one adjacency. The number of them is at most n, thus the number of DCJs applicable for a genome with n genes cannot be more than $4n^2-n$. The length of a most parsimonious DCJ scenario is at most $n-1$, thus the number of most parsimonious DCJ scenarios is a most $(4n^2-n)^{n-1}$. ■

The following lemma sets the largest temperature we need.

LEMMA 4. *If the inverse of the temperature of the Markov chain is greater than $(n-1)\log(4n^2-n)$, then the probability of the HP scenarios is at least $\frac{1}{2}$ in the target distribution.*

PROOF. $c(S)$ is at least 1 for any non-HP path, thus the probability of any non-HP path is at least $(4n^2-n)^{n-1}$ times smaller than that of a HP path. Since there are less than $(4n^2-n)^{n-1}$ times more non-HP paths than HP paths, the probability of the HP paths in the target distribution is at least $\frac{1}{2}$. ■

The following lemma tells what difference between the temperature of neighbour chains is necessary for a swapping probability greater or equal than $\frac{1}{2}$.

LEMMA 5. *If the difference between the inverse temperatures is $\frac{4\log 2}{n(n-2)}$, then the swapping probability given by Equation (7) is at least $\frac{1}{2}$.*

PROOF. Let Δc denote the difference between $c(R_i)$ and $c(R_{i+1})$ and let ΔT denote $\frac{1}{T_{i+1}}-\frac{1}{T_i}$. Equation (7) can be simplified as

$$\min\left\{1, e^{-\Delta c \Delta T}\right\} \tag{11}$$

Since Δc is at most $\frac{n(n-2)}{4}$, the swapping probability is at least $\frac{1}{2}$. ■

B.1. Proof of Theorem 3

We set the temperature of the first chain to infinite, as prescribed, and the temperature of the $i+1^{st}$ chain as

$$T_{i+1} := \frac{1}{\frac{4\log 2}{n(n-2)}+\frac{1}{T_i}} \tag{12}$$

The swapping probability between two chains will be at least $\frac{1}{2}$, based on Lemma 5. The chain with index $\left\lceil \frac{n(n-1)(n-2)\log(4n^2-n)}{4\log 2}+1 \right\rceil$ will have temperature at most $\frac{1}{(n-1)\log(4n^2-n)}$, and thus, the probability of the HP scenarios in its target distribution is at least $\frac{1}{2}$. ■