A Markov chain Monte Carlo approach to reconstructing ancestral genome arrangements

Bret Larget
Department of Mathematics and Computer Science
Duquesne University

Joseph B. Kadane
Department of Statistics
Carnegie Mellon University

Donald L. Simon
Department of Mathematics and Computer Science
Duquesne University

March 6, 2002

Corresponding Author Contact Information:
Bret Larget
Department of Mathematics and Computer Science
Duquesne University
College Hall 440
Pittsburgh, PA, USA, 15282
`larget@mathcs.duq.edu`
(412) 396-6469
(412) 396-5197 (FAX)

Running Title: An MCMC approach to genome arrangements

Key words: Bayesian statistics, gene inversion, genome rearrangement, phylogeny, signed permutations, sorting by reversal

ABSTRACT

We describe a Bayesian approach to infer phylogeny and ancestral genome arrangements on the basis of genome arrangement data using a model in which gene inversion is the sole mechanism of change. A Bayesian approach provides a means to quantify the uncertainty in the phylogeny and in the ancestral genome arrangements. We describe a method of sampling phylogenies from the posterior distribution via Markov chain Monte Carlo (MCMC) that is computationally feasible for large data sets. We compare and contrast this MCMC approach with methods which reconstruct maximum parsimony phylogenies from genome arrangement data and demonstrate several advantages of a Bayesian approach to this problem. Furthermore, we have found that our sampler has discovered many genome rearrangement scenarios that require fewer gene inversions on a Campanulaceae cpDNA data set than other published reconstructions which were thought to be most parsimonious.

1 Introduction

Phylogenetic inference on the basis of molecular sequence data is an active area of research with a long history. Swofford *et al.* (1996) is an excellent review of the field with descriptions of many methods for phylogenetic tree reconstruction including distance-based methods such as neighbor-joining, maximum parsimony, and maximum likelihood, and extensive references to other published work in the field. More recently, Bayesian approaches have been investigated by many authors (Rannala and Yang, 1996; Yang and Rannala, 1997; Mau *et al.*, 1999; Larget and Simon, 1999; Li *et al.*, 2000). Huelsenbeck *et al.* (2001) addresses the recent impact of Bayesian methods on evolutionary biology.

The above methods depend on finding homologous genes from the taxa of interest and aligning the sites accurately. For distantly related taxa, proper alignment of DNA sequences can be highly uncertain and very problematical due to insertions and deletions that change the gene lengths and very long periods of time for multiple nucleotide substitutions to accumulate. Phylogenetic inference methods that jointly handle uncertainty in the alignment as well as the phylogeny are in their infancy and are rather unsatisfactory. Consequently, methods that assume alignment to be accurate exaggerate the certainty of the inferences.

Processes that rearrange entire genomes are thought to be much rarer than processes that affect genetic data at the sequence level, so genome arrangements may be more informative about deep evolutionary relationships than analyses of sequence level data. Recent advances in large-scale sequencing are providing genome arrangement data, spurring efforts to develop methodologies to analyze the data to infer both phylogeny and ancestral genome arrangements.

Genome arrangements are represented abstractly as signed permutations, where each permutation element represents either a gene or a block of genes. Elements of the same sign correspond to genes located on the same strand. Gene inversions are rearrangement events that correspond to reversals of signed permutations, where the reversal changes both the order and the signs of the affected elements. Circular genomes with $n+1$ gene blocks may be represented as signed permutations of length n by choosing an arbitrary reference gene and reading the remaining genes around the circle.

The very simplest type of analysis attempts to reconstruct the genome rearrangement events that separate two genome arrangements. Hannenhalli and Pevzner (1995) found the first polynomial time

algorithm for computing the reversal distance between any two arrangements. Kaplan *et al.* (1999) and Bader *et al.* (2001) simplified and improved it.

There has been more effort put forth recently toward the development of methods to infer phylogeny and ancestral genome arrangements among three or more species. The most studied approach is based on the principle of maximum parsimony: reconstructions that involve the smallest possible number of genome rearrangements are sought. Most parsimonious reconstructions are thought to be the most likely explanations of the true evolutionary past. This framework of analysis begins with the pairwise distance between genome arrangements. This distance can be defined in different ways. The breakpoint distance between two genome arrangements counts the number of adjacent pairs of genes in one arrangement that are not present in the other. This distance is not directly a function of any presumed mechanism for rearrangement. The reversal distance counts the minimal number of gene inversions necessary to transform one arrangement into another. Additional distances could be defined by allowing other types of rearrangement, such as gene transposition. In the present work we restrict consideration to unichromosomal genome arrangements and processes that rearrange genomes on a single chromosome.

Cosner *et al.* (2000b) describes the Maximum Parsimony for Rearranged Genomes Problem as the search for a tree and genome arrangements at the internal nodes to minimize the sum of the pairwise distances over branches of the tree. If the distance measure counts breakpoints, an optimal tree is called a minimum-breakpoint tree. Sankoff and Blanchette (1998) and M. Blanchette (1999) describe a computational method to search for minimum-breakpoint trees. Cosner *et al.* (2000b) and Moret *et al.* (2001) describe subsequent improvements to this approach which increase the speed of finding minimum-breakpoint trees substantially, and also allow searches for most parsimonious trees that minimize the total number of gene inversions. The Multiple Genome Rearrangement Problem (Bourque and Pevzner, 2002) is the same problem in the special case where gene inversions are the only rearrangement mechanism. Solutions to this problem are most parsimonious in that they require the smallest number of total changes, or the smallest number of rearrangement events when the distance measure counts rearrangements.

In previous work, we have approached the problem of phylogenetic inference from genome arrange-

ments from a very different perspective. Simon and Larget (2001) describe a Bayesian approach to the problem that was limited to small simulated data sets. Our recent work (Larget *et al.*, in press) solves the computational difficulties that limited our previous approach and describes a Bayesian method of inference that is computationally tractable for genuine data sets. The types of inference possible in a Bayesian analysis are very different from those made within the maximum parsimony framework. Specifically, our analyses include calculations of uncertainty in both the inferred ancestral sequences and the phylogeny. The remainder of this paper compares a Bayesian approach with maximum parsimony as applied to several example data sets.

2 RESULTS

We first show that arrangements that are closer in reversal distance are not necessarily more likely. Assume that we have a small artificial genome with nine genes arranged in a circle, so the arrangements are represented by signed permutations of size eight. Consider these two examples:

$$p_1 = (8, 3, 7, 1, -5, -4, -6, 2) \quad \text{and} \quad p_2 = (2, 3, 4, 5, 6, 8, 1, 7). \tag{1}$$

The first permutation requires four reversals to sort, the second five. While it might be supposed that the first permutation would be more likely than the second if a random number of random reversals with mean equal to the actual distance of the first permutation from the identity (i.e. four) were applied to the identity permutation, this turns out not to be the case. Applying a Poisson(4) distributed number of random reversals to the identity permutation with all possible reversals being equally likely, the second arrangement is more than twice as likely as the first. The reason is that there is but a single sequence of four reversals that sorts the first permutation while there are 200 sequences of reversals of length five that sort the second. Table 1 contains counts of the number of short sorting sequences for the two permutations.

[Table 1 should appear about here.]

There are a total of 36 possible reversals for permutations of length eight. The probability of achieving these permutations after applying a Poisson(4) distributed number of random reversals to

the identity permutation may be calculated by conditioning on the realized number of reversals.

$$\text{P(identity to } p) = \sum_{k=0}^{\infty} \frac{\text{P(exactly } k \text{ reversals)} \times (\#\text{ of sorting sequences of } p \text{ of length } k)}{(\text{total } \#\text{ of sequences of length } k)} \qquad (2)$$

Truncating this sum at $k = 7$, the probability of p_1 is approximately $2.8 \times \exp(-4) \times 4^4/(36)^4$ while the probability of p_2 is approximately $6.5 \times \exp(-4) \times 4^4/(36)^4$, more than twice as large. This indicates that the most parsimonious reconstructions may not be the most likely, and that methods that account for multiple sorting sequences may be more accurate.

Herpes virus example. Bourque and Pevzner (2002) reanalyzes a small virus data set studied in Hannenhalli *et al.* (1995) with Herpes simplex virus (HSV), Epstein-Barr virus (EBV), and Cytomegalovirus (CMV). The unrooted tree relating these viruses contains a single internal ancestral node with edges to each of the three leaves. Hannenhalli *et al.* (1995) reduce the gene arrangements to signed permutations of seven gene blocks and find two most parsimonious rearrangement scenarios that each require seven total rearrangements. Bourque and Pevzner (2002) do not block the genes with common arrangements in the three viruses, and analyze three signed permutations of length 25, finding a single rearrangement scenario with seven total rearrangements.

In our analysis, we use Updates 1 and 2 from the Markov chain Monte Carlo method described in Larget *et al.* (in press) to sample the possible rearrangement scenarios. Under a model we describe below in the Methods section, we are able to compute the posterior distribution of the ancestral sequence. Figure 1 shows these results. In this example, there is an 86 percent probability that the true rearrangement history is one of the most parsimonious reconstructions with seven total rearrangements. Additionally, the two possible ancestral arrangements in these most parsimonious reconstructions have a combined posterior probability of 90 percent because they can also occur in reconstructions with more than seven total events. Six different ancestral arrangements account for nearly 99 percent of the posterior probability.

[Figure 1 should appear about here.]

Human, Fruit Fly, and Sea Urchin mitochondrial arrangements. Sankoff *et al.* (1996) and Bourque and Pevzner (2002) analyze the mitochondrial genome arrangements of human, sea urchin,

and fruit fly. We use the full mitochondrial arrangements in Boore (2001) with 37 genes leading to signed permutations of length 36. The other authors blocked some genes to find shorter permutations of length 33. Bourque and Pevzner (2002) report a single most parsimonious reconstruction that requires a total of 39 reversals. We find that there are at least 80 unique ancestral arrangements consistent with the most parsimonious reconstructions (and there are may be more). Figure 2 shows the posterior distribution for the total number of gene inversions. Interestingly, we calculate the posterior probability that the true tree has a most parsimonious reconstruction scenario to be only about one percent. Without restriction to most parsimonious reconstructions, we need nearly 7000 different ancestral arrangements to account for 90 percent of the posterior probability.

[Figure 2 should appear about here.]

Campanulaceae chloroplast genome arrangements. Cosner *et al.* (2000a,b), Moret *et al.* (2001), and Bourque and Pevzner (2002) analyze a data set of chloroplast genome arrangements with 105 markers from twelve Campanulaceae genera plus the outgroup tobacco. These arrangements are in Table 2. (Cosner *et al.* (2000a) and Cosner *et al.* (2000b) contain several typographical errors in reporting these genome arrangements. The arrangements in Table 2 are consistent with the data set available on the Web site of one of the authors of these papers.)

[Table 2 should appear about here.]

This data problem is more complicated than the previous examples because there is considerable uncertainty in the true phylogeny as well as in the ancestral arrangements. For thirteen taxa, there are 13,749,310,575 possible unrooted binary trees. Based on a heuristic search of part of the tree space, Moret *et al.* (2001) reports 216 most parsimonious trees, each of which requires 67 total gene inversions. Using a different heuristic search method, Bourque and Pevzner (2002) reports a single tree with 65 total gene inversions indicating that Moret *et al.* (2001) was incorrect.

In our analysis, we find the posterior distribution of the tree topology to be rather dispersed. The most likely tree topologies, of which there are several hundred, have posterior probabilities of approximately 0.4 percent each. We need over 180 tree topologies to account for 50 percent of the

posterior probability, over 390 to account for 90 percent, over 420 to account for 95 percent, and over 480 to account for 99 percent. Summarizing this uncertainty with a single tree topology is inadequate. Furthermore, the posterior probability that the true tree has a most parsimonious reconstruction is only 22 percent. The distribution is: 64 reversals, 22 percent; 65 reversals, 61 percent; 66 reversals, 13 percent; 67 or more reversals, 3 percent. (Percentages do not sum to 100 percent because of round-off error.)

We find 180 different tree topologies that require only 64 inversions, fewer than the 65 inversions in Bourque and Pevzner (2002) and the 67 in Moret *et al.* (2001). These 180 tree topologies have similar structure. Labels in the following description follow Table 2. In each case, the following subclades appear and are grouped as a single clade we label X: 1, [2,3], 4, and [8,9]. The fifteen possible rooted tree topologies of these four subclades each appear the same number of times in the 180 trees. The clade we label Y includes X, 5, 6, and 7 is in all 180 tree topologies. Only four of the fifteen possible tree topologies are present in equal number: ((X,5),(6,7)); (((X,5),7),6); (((5,7),X),6); and (((5,7),6),X). Finally, there are three possible ways to combine the clades X, Y, and [10,11]. These each appear the same number of times. Each of the 180 tree topologies we found that require only 64 inversions is characterized by the choices from among the 15, 4, and 3 subtree topologies respectively. Our sampler found no other tree topologies that required 64 or fewer inversions.

3 DISCUSSION

The best case for maximum parsimony methods is in the case when the most parsimonious reconstruction is very likely to be correct. Then a biologist interpreting the results has a good basis with which to start. For example, in the herpes virus example, one ancestral arrangement has a substantial amount of posterior probability and is not too bad of a summary by itself. But if individual most parsimonious reconstructions are very unlikely, there is a high degree of uncertainty about which reconstruction is correct. In the human, fruit fly, sea urchin example, there is tremendous uncertainty in the ancestral arrangement. We found 80 different ancestral arrangements consistent with most parsimonious reconstructions and their combined posterior probability is small. Furthermore, in this example it is very likely that the true rearrangement scenario is not one of the most parsimonious reconstructions.

To report a single ancestral arrangement in this case is highly misleading. The real difficulty is that maximum parsimony methods provide no warning when the single reconstruction selected has low probability of being correct.

By contrast, Bayesian methods report a full posterior distribution on the space of possible trees. If one of those is very likely (whether it is most parsimonious or not), that fact will be evident from the distribution. If there are many, roughly equally likely trees or ancestral arrangements, that also will be evident. We submit that Bayesian analyses are more likely to be useful to biologists than are maximum parsimony methods because accounting for and quantifying uncertainty is important.

The Bayesian analyses have other virtues as well. Because the Markov chain Monte Carlo sampler spends the bulk of its time on trees of high probability, it coincidentally can find better maximum parsimony trees than found by other computational approaches for some data sets. For example, in the Campanulaceae data set, we found trees with 64 inversions, while programs searching for trees with few inversions reported no better than 65 inversions. We expect that other researchers interested in finding most parsimonious reconstructions may find stochastic search based on MCMC to be more efficient than current heuristic optimization methods. Bourque and Pevzner (2002) describe the Campanulaceae data set with its 13 taxa as "one of the most challenging genome rearrangement data sets". Larget *et al.* (in press) successfully applies the Bayesian approach used in this paper to a data set with 19 taxa, a problem in which the tree space is more than 460 million times as large.

A Bayesian approach has other benefits. First, it is possible to incorporate gracefully other sources of information. This information may come from previous studies on other data. Furthermore, it is straightforward in principle to extend our current model by adding other mechanisms of genome rearrangement or to use prior information about inversion hot spots to remove the assumption that all possible inversions are equally likely. Extending the approach to the multichromosomal data sets described in Bourque and Pevzner (2002) should also be possible.

4 METHODS

We assume a very simple model of genome rearrangement, with gene inversion as the sole mechanism. We assume that the evolutionary relationships among the taxa in our analysis are described by a

phylogeny in which each speciation event results in two lineages. We do not assume a molecular clock, so the overall rate of gene inversion may be different for different lineages. Our prior distribution is that all unrooted tree topologies are equally likely. Branches of the unrooted tree have independent lengths selected from a Gamma distribution. Given a branch length, a Poisson number of gene inversions with this mean are realized. Their locations on the branch are independent and uniformly distributed at random. Given that a gene inversion occurs, we assume that all possible gene inversions are equally likely. The likelihood of the observed data conditional on the tree topology, branch lengths, and inversion scenario, is simply an indicator function that the data is consistent with the tree topology and the order of the specific gene inversion events on each branch. We are able to integrate out analytically the specific dependence on the branch lengths and the absolute locations of the gene inversions.

The state space for our Markov chain consists of the tree topology, the gene inversion counts on each branch, and the relative order in which the specific inversions occur constrained to be consistent with the observed arrangments. We propose changes in this state space by cycling through three different updates. The first update changes the inversion scenario on the three branches adjacent to an internal node. The second update changes the inversion scenario on a single branch. The third changes the tree topology and modifies the inversion scenario on two of the affected branches. Larget *et al.* (in press) contains a full description of the computational details of this approach.

5 ACKNOWLEDGEMENTS

Bret Larget thanks the Department of Statistics at Carnegie Mellon University where he was a visitor during the writing of this paper. This research is supported in part by National Science Foundation grant DEB-0075406.

References

Bader, D. A., B. M. E. Moret, and M. Yan 2001 A linear-time algorithm for computing inversion distances between signed permutations with an experimental study. Journal Computational Biology **8**: 483–491.

Boore, J. 2001 Mitochondrial gene arrangement source guide, version 6.0. http://www.jgi.doe.gov/Mitochondrial_Genomics.html, DOE Joint Genome Institute.

Bourque, G. and P. Pevzner 2002 Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Research **12**: 26–36.

Cosner, M. E., R. K. Jansen, B. M. E. Moret, L. A. Raubeson, L.-S. Wang, T. Warnow, and S. Wyman 2000a An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In *Comparative Genomics (DCAF-2000)*, Kluwer Academic Publishers, Montreal, Canada.

Cosner, M. E., R. K. Jansen, B. M. E. Moret, L. A. Raubeson, L.-S. Wang, T. Warnow, and S. Wyman 2000b A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB-2000)*, AAAI Press, Menlo Park, California.

Hannenhalli, S., C. Chappey, E. Koonin, and P. Pevzner 1995 Genome sequence comparison and scenarios for gene rearrangments: A test case. Genomics **30**: 299–311.

Hannenhalli, S. and P. Pevzner 1995 Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proceedings of the Twenty-seventh annual ACM-SIAM symposium on the theory of computing*, ACM Press, New York.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. Bollback 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**: 2310–2314.

Kaplan, H., R. Shamir, and R. Tarjan 1999 Faster and simpler algorithm for sorting signed permutations by reversals. SIAM Journal on Computing **29**: 880–892.

Larget, B. and D. L. Simon 1999 Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution **16**: 750–759.

Larget, B., D. L. Simon, and J. B. Kadane in press Bayesian phylogenetic inference from animal mitochondrial genome arrangements. Journal of the Royal Statistical Society, Series B .

Li, S., H. Doss, and D. Pearl 2000 Phylogenetic tree reconstruction using Markov chain Monte Carlo. Journal of the American Statisical Society **95**: 493–508.

M. Blanchette, D. S., T. Kunisawa 1999 Gene order breakpoint evidence in animal mitochondrial phylogeny. Journal of Molecular Evolution **49**: 193–203.

Mau, B., M. A. Newton, and B. Larget 1999 Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**: 1–12.

Moret, B. M. E., L. Wang, T. Warnow, and S. Wyman 2001 New approaches for reconstructing phylogenies from gene order data. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB-2001)*.

Rannala, B. and Z. Yang 1996 Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution **43**: 304–311.

Sankoff, D. and M. Blanchette 1998 Multiple genome rearrangement and breakpoint phylogeny. Journal of Computational Biology **5**: 555–570.

Sankoff, D., G. Sundaram, and J. Kececioglu 1996 Steiner points in the space of genome rearrangements. International Journal of the Foundation of Computer Science **7**: 1–9.

Simon, D. L. and B. Larget 2001 Phylogenetic inference from mitochondrial genome arrangement data. In *Computational Science — ICCS 2001*, edited by V. Alexandrov, J. Dongarra, B. Juliano, R. Renner, and C. Tan, number 2074 in Lecture Notes in Computer Science, Springer-Verlag.

Swofford, D., G. Olsen, P. Waddell, and D. Hillis 1996 Phylogenetic inference. In *Molecular Systematics*, edited by D. M. Hillis, C. Moritz, and B. K. Mable, Sinauer Associates.

Wyman, S. 2000 http://www.cs.utexas.edu/users/stacia/ismb2000/, Department of Computer Science, University of Texas, Austin.

Yang, Z. and B. Rannala 1997 Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution **14**: 717–724.

Herpes virus data

Virus	Arrangement
HSV:	(1–16) ($\overline{19–17}$) (20–23) ($\overline{25–24}$)
EBV:	(1–16) ($\overline{20–17}$) (21–25)
CMV:	(1–11) ($\overline{13–12}$) ($\overline{16–14}$) ($\overline{25–24}$) (17–23)

Ancestral arrangements

Label	Prob.	Cum. Prob.	Arrangement
A1:	0.627	0.627	(1–25)
A2:	0.276	0.903	(1–23) ($\overline{25–24}$)
A3:	0.039	0.943	(1–16) ($\overline{20–17}$) (21–25)
A4:	0.033	0.975	(1–16) ($\overline{19–17}$) (21–23) ($\overline{25–24}$)
A5:	0.007	0.983	(1–16) ($\overline{20–17}$) (21–23) ($\overline{25–24}$)
A6:	0.007	0.989	(1–16) ($\overline{19–17}$) (21–25)

Posterior distribution of total number of inversions

Number of inversions	Probability
7	0.861
8	0.104
9	0.032
10+	0.003

Figure 1: The first table shows the genome arrangements of the three viruses. The second table contains a summary of the posterior distribution on the space of possible arrangements for the ancestral node. Nearly 99 percent of the posterior probability is concentrated on only six of the $2^{25}25!$ possible arrangements.

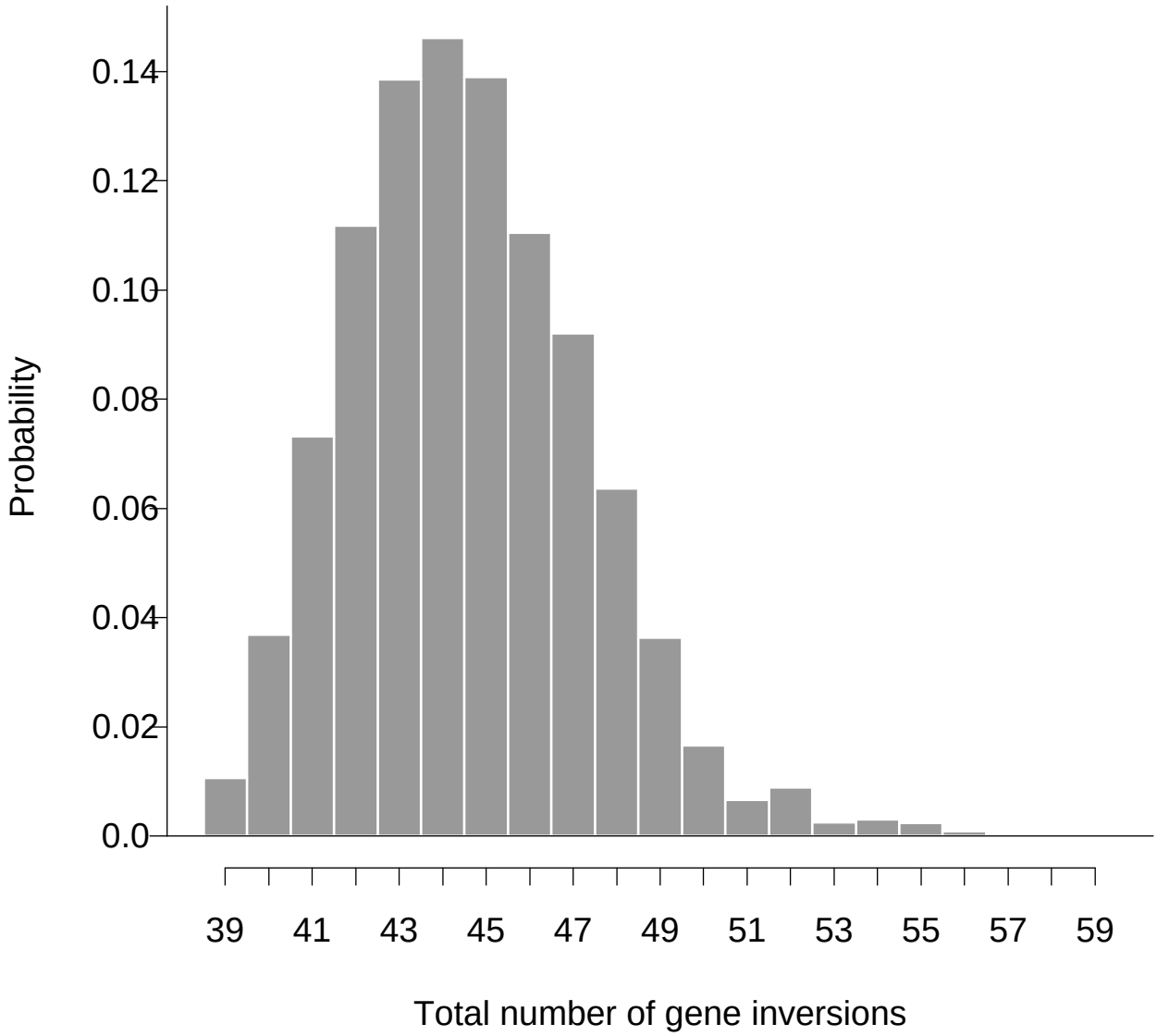

Figure 2: **Histogram of the probability distribution of the total number of inversions in the human, fruit fly, sea urchin example.**

		# of sorting sequences			
Permutation	Distance	4	5	6	7
p_1	4	1	8	791	9,918
p_2	5	0	200	2,668	147,282

Table 1: **Numbers of sorting sequences.** For the signed permutations in Equation 1, the second column lists the minimal number of reversals to sort, and the remaining columns contain the number of distinct sorting sequences by length.

Label	Genera	Arrangement
1	Trachelium	(1–15) ($\overline{76–56}$) ($\overline{53–49}$) (37–40) ($\overline{35–26}$) ($\overline{44–41}$) (45–48) ($\overline{36}$) ($\overline{25–16}$) ($\overline{90–84}$) (77–83) (91–96) ($\overline{55–54}$) ($\overline{105–97}$)
2	Campanula	(1–15) ($\overline{76–56}$) ($\overline{53–49}$) ($\overline{39–37}$) (40) ($\overline{35–26}$) ($\overline{44–41}$) (45–48) ($\overline{36}$) ($\overline{25–16}$) ($\overline{90–84}$) (77–83) (91–96) ($\overline{55–54}$) ($\overline{105–97}$)
3	Adenophora	(1–15) ($\overline{76–56}$) ($\overline{53–49}$) ($\overline{39–37}$) (28–35) (40) (26–27) ($\overline{44–41}$) (45–48) ($\overline{36}$) ($\overline{25–16}$) ($\overline{90–84}$) (77–83) (91–96) ($\overline{55–54}$) ($\overline{105–97}$)
4	Symphyandra	(1–15) ($\overline{76–56}$) ($\overline{39–37}$) (49–53) (40) ($\overline{35–26}$) ($\overline{44–41}$) (45–48) ($\overline{36}$) ($\overline{25–16}$) ($\overline{90–84}$) (77–83) (91–96) ($\overline{55–54}$) ($\overline{105–97}$)
5	Legousia	(1–4) (9–15) ($\overline{76–56}$) ($\overline{27–26}$) ($\overline{44–41}$) (45–48) ($\overline{36–35}$) ($\overline{25–16}$) ($\overline{90–84}$) (77–83) (91–96) (5–8) ($\overline{55–53}$) ($\overline{105–98}$) (28–34) ($\overline{40–37}$) (49–52) ($\overline{97}$)
6	Asyneuma	(1–15) ($\overline{76–61}$) ($\overline{56–53}$) ($\overline{60–57}$) ($\overline{27–26}$) ($\overline{44–41}$) (45–48) ($\overline{36–35}$) ($\overline{25–16}$) ($\overline{89–84}$) (77–83) (90–96) ($\overline{105–98}$) (28–34) ($\overline{40–37}$) (49–52) ($\overline{97}$)
7	Triodanus	(1–15) ($\overline{76–56}$) ($\overline{27–26}$) ($\overline{44–41}$) (45–48) ($\overline{36–35}$) ($\overline{25–16}$) ($\overline{89–84}$) (77–83) (90–96) ($\overline{55–53}$) ($\overline{105–98}$) (28–34) ($\overline{40–37}$) (49–52) ($\overline{97}$)
8	Wahlenbergia	(1–11) ($\overline{60–56}$) ($\overline{53–49}$) (37–40) ($\overline{35–28}$) (12–15) ($\overline{76–61}$) ($\overline{27–26}$) ($\overline{44–41}$) (45–48) ($\overline{36}$) (54) ($\overline{25–16}$) ($\overline{90–84}$) (77–83) (91–96) ($\overline{55}$) ($\overline{105–97}$)
9	Merciera	(1–10) (49–53) (28–35) ($\overline{40–37}$) ($\overline{60–56}$) (11–15) ($\overline{76–61}$) ($\overline{27–26}$) ($\overline{44–41}$) (45–48) ($\overline{36}$) (54) ($\overline{25–16}$) ($\overline{90–85}$) (77–84) (91–96) ($\overline{55}$) ($\overline{105–97}$)
10	Codonopsis	(1–8) ($\overline{36–18}$) ($\overline{15–9}$) (40) (56–60) (37–39) ($\overline{44–41}$) (45–53) (16–17) (54–55) (61–76) ($\overline{96–77}$) ($\overline{105–97}$)
11	Cyananthus	(1–8) (28) ($\overline{36–29}$) ($\overline{27–26}$) (40) (56–60) (37–39) ($\overline{25–9}$) ($\overline{44–41}$) (45–48) ($\overline{55–49}$) (61–96) ($\overline{105–97}$)
12	Platycodon	(1) (8) (2–5) (29–36) ($\overline{56–50}$) ($\overline{28–26}$) (9) ($\overline{49–45}$) (41–44) (37–40) (16–25) (10–15) (57–59) (6–7) (60–96) ($\overline{105–97}$)
13	Tobacco	(1–105)

Table 2: **Campanulaceae arrangements.** Chloroplast genome arrangements of twelve genera of Campanulaceae and the outgroup tobacco are displayed in maximal gene blocks relative to the outgroup tobacco. The notation (37–40) stands for the sequence $37, 38, 39, 40$ while the notation ($\overline{44–41}$) represents the sequence $-44, -43, -42, -41$ with similar notation for single genes. This data is at `http://www.cs.utexas.edu/users/stacia/ismb2000/`.