

---

# Comparative Genomics

---

**J. Dicks and G. Savva**

*John Innes Centre, Norwich, UK*

Over the last couple of decades, numerous genetic and physical maps have been developed for a wide range of species. This has led to the field of *comparative genomics* which analyses characteristics of whole genomes, in contrast to the analysis of single genes in comparative genetics. Where markers on one map are known to have homologues on another, we can get a feel for the relative distribution of markers on the chromosomes of different species. From these we can deduce segments of chromosomes where the gene content, and sometimes also the order of the genes, is similar in two or more species. More recently, we have begun to see the DNA sequences of whole genomes, which enables us to compare genomes at both the sequence and gene levels. The promise of the post-genome era means that we will see many fully sequenced genomes within the next decade, and we should develop analytical methods that are mature when significant quantities of data become available. However, not all species will be sequenced and so we should continue to develop methods of analysis that are appropriate for both types of data. In this chapter, we will give a flavour of the types of comparative genome analysis that are carried out. We will highlight particular problems, such as phylogenetic inference and the use of maps to compare genomes. Finally, we will give examples of problems with our current analyses that could be solved in the future.

## 5.1 INTRODUCTION

Every eukaryotic organism possesses a *nuclear genome* (so called because it is found in the nucleus of every cell) containing one or more linear chromosomes. Additionally, each organism will possess one or more of a small number of linear supernumerary (or B) chromosomes, a small circular mitochondrial genome and a small circular chloroplast genome. The data sets most widely used for comparative genome analysis are currently nuclear and mitochondrial genomes. Mitochondria are well characterized for many organisms due to their small size, contain few genes and are relatively easy to analyse. Nuclear genomes are less well characterized for fewer organisms, contain a much greater number of genes and are much more difficult to analyse. However, nuclear genomes

contain a great deal more information than mitochondrial genomes, including most of the genes involved in species improvement and disease, and therefore we should be able to analyse this more difficult case.

Each chromosome contains a single *centromere*, sometimes known as the *waist* of the chromosome. This plays an important role in cell division, and we will see below that it is also involved in some chromosomal mutations. The position of the centromere varies from the 'top' of the chromosome to the middle. Genes are arranged linearly on the chromosomes, with adjacent genes separated by *intergenic DNA*. Intergenic DNA is often called junk DNA, and the amount varies enormously between species. However, it is thought to contain elements of functional and evolutionary significance and therefore should not be ignored completely when comparing genomes. For the rest of this chapter will speak of comparing the genomes of species. However, the methods outlined here could equally be used to analyse the genomes of other taxonomic groups such as subspecies.

We would like to compare the genomes of two or more species. This basic statement covers a multitude of data types. Most comparative analyses focus on *gene homologues*. Two genes that are both related by descent from a gene in a common ancestral species are said to be homologous to one another. Close homologues usually share a similar DNA sequence and biological function, although a similar sequence does not necessarily imply a similar function. When we look at the *gene content* of two genomes we are looking to see how many homologues they share and how many genes each contains for which the other does not have a homologue. We would expect closely related species to have roughly the same gene content but for more distantly related species to contain species- or group-specific genes. Also we might wish to examine the *gene orders*, literally the order in which the genes are found upon the chromosomes of their respective species.

There are many other ways to compare genomes. We may be interested in functional data and compare the protein structures or functions of the proteins encoded by the genes. We may not be interested in genes at all and prefer to look at intergenic entities such as *transposable elements*. Given time, we will develop methodologies to look at all of these data types, to compare the results of the different methods and perhaps even to integrate the results. For now, many researchers are looking at problems posed by *gene order* and *gene content*, and in this chapter we will focus on the issues they raise.

What can we learn from the gene content and gene order of an organism? There are essentially two types of analysis that we might like to perform: *comparative mapping* and *evolutionary*. A comparative mapping analysis can help us answer questions such as:

- Where in species 2 is the homologue of gene *A* from species 1?
- What is the likely function of gene *B* in species 3?
- How many chromosomal segments would we see if we knew about every gene within a genome, rather than the subset we see within a map?

An evolutionary analysis may help us:

- estimate a chromosomal evolutionary tree;
- hypothesize the gene order and content of an ancestral species;

- find the sequence of mutations that could have changed the gene order of one genome to that of another;
- find the most likely sorts of mutation events that gave rise to a certain set of species.

In Section 5.2 we will look at comparative mapping and the problems that comparing maps can solve. We will define the complex nature of chromosomal evolution in Section 5.3, and we will use the concepts described by it to look at measures of genome difference in Section 5.4. We will introduce models of chromosomal evolution in Section 5.5, and we show in Section 5.6 how these models, together with the measures of genome difference, can be used to derive an evolutionary tree based on chromosomal, rather than genic, evolution. In Section 5.7 we will note briefly the considerable efforts in the computer science community to tackle similar problems to these, and we will introduce software to carry out comparative genome analysis in Section 5.8. Finally, in Section 5.9, we will look at areas of potential future research. In addition to the overview given here, Sankoff and Nadeau (2000) give a thorough introduction to biological, statistical and computational advances in comparative genomics.

## 5.2 COMPARATIVE MAPPING

A comparative mapping analysis looks at the maps of more than one species, together with information about homologous relationships between the mapped markers. Maps can be used to extrapolate information from one species, on the location and function of markers and genes, to be applied to another. For example, imagine that we know about gene *A*, which is involved in an important function in species 1. We would like to know if there is a ‘similar’ gene in species 2, that perhaps carries out a similar function. We know that gene *A* is flanked by genes *B* and *C* in species 1. If we know the locations of the homologues of genes *B* and *C* in species 2 we can predict the location of a potential homologue in species 2. This provides a starting point for further experimentation to confirm or refute this prediction. This problem is often made more difficult by gene duplication, where we cannot always know if two genes are *orthologous* (identical by descent) or *paralogous* (identical by duplication) and we may look for a gene in the wrong part of the genome. Alternatively, we may be interested in gene *D* in species 3 but we do not know its function. By looking at the functions of its homologues in other species and perhaps the functions of its close neighbours, we may gain more information on a putative function.

When comparing genomes by their maps, we have information about homologous markers but we have gaps between those markers and we do not know how many genes lie within those gaps or where their homologues lie in other species. Other data sets may have even less information, where a marker is known to be located on a particular chromosome (a *chromosome assignment*) but its map position is not known. This ‘missing’ information will undoubtedly bias the results of any analysis carried out on those data. Despite this, it is possible to make inferences about chromosomal segments and evolution. Nadeau and Taylor (1984) used mapping data from the autosomal chromosomes of human and mouse to estimate the number and lengths of conserved chromosomal segments between these two species. They showed that the length of a segment  $\hat{m}$ , given the assumption that

markers and breakpoints were scattered randomly across a genome (the *random breakage model*), could be found as follows:

$$\hat{m} = \frac{r(n+1)}{n-1},$$

where  $n$  is the number of markers in a segment and  $r$  is the distance between the two outermost markers. They reasoned that segments with fewer than two markers would be missed, and they used this to derive the expected sample mean of the transformed segment lengths  $\hat{m}$ :

$$E[x'] \cong \frac{L^2D + 3L}{LD + 1},$$

where  $L$  is the mean length of conserved segments measured in centimorgans and  $D$  is the density of mapped homologous loci in the genome. They further estimated the number of homology disruptions between human and mouse.

Waddington (2000) and Waddington *et al.* (2000) generalized the work of Nadeau and Taylor in order to compare maps of chickens with those of human and mouse. They no longer assumed that chromosome lengths were large when compared to segment lengths (important as chicken linkage groups vary considerably in length, with some likely to contain just one segment while others may contain several) and they did not assume chromosome breakage occurred at random. They used the beta distribution to model segment lengths, and used a scaling parameter  $l_k$ , the length of chromosome  $k$ , to derive the density function of segment length  $y$  on chromosome  $k$ :

$$f(y_k) = l_k^{-1} \left(\frac{y}{l_k}\right)^{a-1} \left(1 - \frac{y}{l_k}\right)^{b-1} \{B(a, b)\}^{-1}$$

where  $a$  and  $b$  are the beta distribution parameters and  $B(a, b)$  the beta function  $\int_0^1 x^{a-1}(1-x)^{b-1} dx$ . Consequently, the expected number of segments on chromosome  $k$  is  $S_k = (a+b)/a$  and the mean segment length on chromosome  $k$  is  $l_k(a+b)/a$ . The special case  $a = 1$  corresponds to the random breakage model. They also derived the distribution of the number of genes  $n$  in a syntenic group (count data), based on a conserved segment of length  $y$ , and the joint distribution of  $n$  and the distance between the two outermost markers in the syntenic group (range data), and used these singly and in combination to analyse the chicken data set. They noted that the flexibility of the new models made them suitable for the analysis of a wide range of data sets, particularly as their formulation enabled testing of the models and data assumptions.

Ehrlich *et al.* (1997) developed methods to estimate *synteny conservation* (i.e. where two or more homologous markers are located on the same chromosome in two or more species). They used their methods to find the percentage of conserved syntenies in human and mouse that had already been observed. They also showed that rates of chromosomal rearrangement varied considerably between mammalian lineages and that inter-chromosomal mutations had occurred four times more often than intra-chromosomal mutations in the lineages leading to human and mouse, despite the strong selective forces against them.

Sankoff *et al.* (2000b) analysed the Nadeau and Taylor approach in light of the large number of genes mapped in both human and mouse since 1984 (approximately 1500 genes, compared to the 83 analysed by Nadeau and Taylor). They concluded that the

results of the Nadeau and Taylor analysis, which are consistent with current evolutionary estimates, were accurate due to the formulation of their model and the data used rather than luck alone. They also looked at the effect of using chromosomal assignment only on the estimation of the number of conserved segments. They noted that, for example, small intra-chromosomal disruptions (e.g. inversions) could result in two chromosomal segments being regarded as one. For this case, they added a correction to their probability function  $P(a, m, n)$ ,

$$P(a, m, n) = \frac{\binom{m-1}{a-1} \binom{n+1}{a}}{\binom{n+m}{m}},$$

the probability of observing  $a$  non-empty segments for  $m$  genes and  $n$  breakpoints, so that

$$Q(b, m, n, c) = \sum_{a=b}^{n+1} P(a, m, n) \binom{a-1}{a-b} \left(\frac{1}{c}\right)^{a-b} \left(\frac{c-1}{c}\right)^{b-1}$$

is the probability that only  $b$  of the  $a$  non-empty segments are counted on the  $c$  chromosomes being examined. They also calculated the number of segments due to inter- rather than intra-chromosomal rearrangements.

Schoen (2000) looked at the effect of marker density on the estimation of the number of chromosomal breakpoints between pairs of species, for the random breakage model. He showed that the estimated number of breakpoints was close to the expected value when marker density was high. However, the amount of rearrangement could be underestimated when marker density was low, particularly when the species being analysed were distantly related. He also showed that underestimation could occur when inversions were common in the divergence of the species and discussed the results of Ehrlich *et al.* (1997) in light of the bias against detecting inversion events.

## 5.3 CHROMOSOMAL EVOLUTION

We know that DNA sequences evolve via the mutation events substitution, insertion and deletion. Similarly, chromosomes evolve via mutations known as *chromosomal rearrangements* that are less frequent but usually much larger than DNA sequence mutations. We consider two types of rearrangement: those that alter the gene content of the chromosome (the *non-conservative* rearrangements) and those that do not (the *conservative* rearrangements). It is also helpful to consider separately the case of a single chromosome evolving via chromosomal rearrangement and that of a set of chromosomes (the multi-chromosome case).

### 5.3.1 Single-chromosome Mutations

Consider a single chromosome with a gene order  $G$ . The chromosome may be circular (mitochondrion or chloroplast) or linear (nuclear genomes consisting of one chromosome or a single chromosome from a larger nuclear genome, where that chromosome evolves conservatively). The representation we will use below considers a linear chromosome. However, this can be adapted easily for the circular chromosome case. The chromosome contains  $N$  genes  $g_i$  (for  $i = 1, \dots, N$ ) that are represented as signed integers, such that

homologues in other species share the same number. The sign of the number represents the orientation of the gene, denoting the way in which it is *transcribed* or read. For example, if we denote a gene as '+1' on our first species then its homologues in all other species must be denoted as '+1' or '-1', depending on their orientations and regardless of their positions on their respective chromosomes. The gene order is therefore simply a signed permutation representing the position and orientation of the genes on each chromosome. Therefore we can write a gene order as follows:

$$G = (g_1, g_2, \dots, g_{N-1}, g_N).$$

For the single-chromosome case, where we constrain the outcome of any mutation to be another single chromosome, the conservative events are *inversion*, intra-chromosomal translocation or *shift*, and inverted intra-chromosomal translocation or *inverted shift*. For computational simplicity we will define all mutation events as the outcomes of breaks *between genes*, in the intergenic regions. We denote the breakpoints  $\tilde{b}$  (from  $0 \dots N$ ) such that breakpoint  $b_i$  lies between genes  $g_i$  and  $g_{i+1}$ . Breakpoints  $b_0$  and  $b_N$  are special breaks that lie between the outermost genes and the *telomeres* or ends of the chromosomes.

An inversion (often termed a *reversal* in the computer science literature) is the result of two breaks, with the central segment turning about  $180^\circ$  before rejoining with the two broken ends. We define the gene order  $G'$  following an inversion  $INV(b_1, b_2)$  between breakpoints  $b_1$  and  $b_2$ , where  $0 \leq b_1 < b_2 \leq N$ :

$$G' = (g_1, \dots, g_{b_1}, -g_{b_2}, \dots, -g_{b_1+1}, g_{b_2+1}, \dots, g_N).$$

A shift is a three-break mutation, where a segment is moved to another part of the same chromosome. We define the gene order  $G'$  following a shift  $SFT(b_1, b_2, b_3)$  between breakpoints  $b_1$ ,  $b_2$  and  $b_3$ , where  $0 \leq b_1 < b_2 < b_3 \leq N$ :

$$G' = (g_1, \dots, g_{b_1}, g_{b_2+1}, \dots, g_{b_3}, g_{b_1+1}, \dots, g_{b_2}, g_{b_3+1}, \dots, g_N).$$

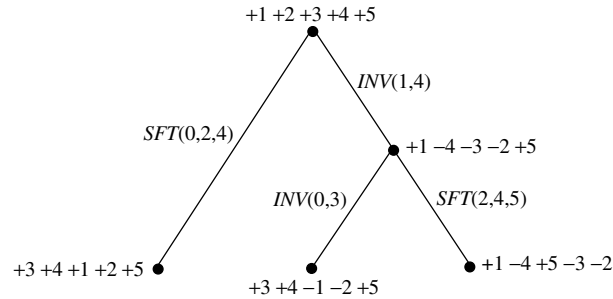
An inverted shift is similar to a shift, but here the segment inverts before it is inserted into its new location. We define the gene order  $G'$  following an inverted shift  $INVSFT(b_1, b_2, b_3)$  between breakpoints  $b_1$ ,  $b_2$  and  $b_3$ , where  $0 \leq b_1 < b_2 < b_3 \leq N$ :

$$G' = (g_1, \dots, g_{b_1}, g_{b_2+1}, \dots, g_{b_3}, -g_{b_2}, \dots, -g_{b_1+1}, g_{b_3+1}, \dots, g_N).$$

Figure 5.1 shows a simple example of a gene order phylogenetic tree, showing changes in gene order via inversions and shifts. In real situations, usually we will not know the states at the tree root and the internal nodes.

The single-chromosome non-conservative events are *tandem duplication* and *deficiency*. A tandem duplication causes a short chromosomal segment to be duplicated and inserted adjacent to the original copy. The new segment can remain in its original position or it can be moved to another location by subsequent mutations. For example, the evolution of the haemoglobin locus, from a single gene to a four-gene complex, is thought to have undergone repeated tandem duplication. We define the gene order  $G'$  following a tandem duplication  $TAN(b_1, b_2)$  between breakpoints  $b_1$  and  $b_2$ , where  $0 \leq b_1 < b_2 \leq N$ :

$$G' = (g_1, \dots, g_{b_1}, g_{b_1+1}, \dots, g_{b_2}, g_{b_1+1}, \dots, g_{b_2}, g_{b_2+1}, \dots, g_N).$$



**Figure 5.1** An example of a gene order tree, for fixed gene content.

A deficiency causes a segment to be removed from a chromosome. We define the gene order  $G'$  following a deficiency  $DEF(b_1, b_2)$  between breakpoints  $b_1$  and  $b_2$ , where  $0 \leq b_1 < b_2 \leq N$ :

$$G' = (g_1, \dots, g_{b_1}, g_{b_2+1}, \dots, g_N).$$

The first five rows of Table 5.1 (with  $C = 1$ ) shows the potential number of distinct events for a chromosome of  $N$  genes. The dagger marks the fact that we may wish, computationally, to limit the size of a tandem duplication. Note that, for a single chromosome of fixed size  $N$  evolving conservatively, the total number of mutations is also fixed. This is not the case for the non-conservative single chromosome, where both  $N$  and the total number of events may vary or for the multi-chromosome case (see below), even for a fixed total gene size of  $N$ .

### 5.3.2 Multi-chromosome Mutations

Here we consider a nuclear genome consisting of two or more chromosomes. Each chromosome  $i$  ( $i = 1$  to  $C$ ) contains  $N_i$  genes and has the gene order  $G_i$ . We denote genes as we did for single chromosomes, such that homologues share the same gene number and that each gene has a positive or a negative orientation. For the multi-chromosome case, any one of its chromosomes may undergo any of the mutations seen for the single-chromosome case. In addition to the single-chromosome mutations, there are other chromosomal rearrangements that alter one or more chromosomes. The additional conservative events are *inter-chromosomal translocation*, *inverted inter-chromosomal translocation*, *reciprocal translocation*, *centric fusion* and *dissociation*. Table 5.1 shows the number of mutations that can occur within a multi-chromosome genome.

An inter-chromosomal translocation (or *insertional translocation*) is a three-break mutation, where a segment from one chromosome becomes inserted at a new location on a second chromosome. We define the gene orders following an inter-chromosomal translocation  $INS(c_1, b_1, b_2; c_2, b_3)$  between breakpoints  $b_1$  and  $b_2$  on chromosome  $c_1$  and breakpoint  $b_3$  on chromosome  $c_2$ , where  $0 \leq b_1 < b_2 \leq N_{c_1}$ ,  $0 \leq b_3 \leq N_{c_2}$ ,  $1 \leq c_1, c_2 \leq C$  and  $c_1 \neq c_2$ , to be:

$$G'_{c_1} = (g_{1,c_1}, \dots, g_{b_1,c_1}, g_{b_2+1,c_1}, \dots, g_{N_{c_1},c_1}),$$

$$G'_{c_2} = (g_{1,c_2}, \dots, g_{b_3,c_2}, g_{b_1+1,c_1}, \dots, g_{b_2,c_1}, g_{b_3+1,c_2}, \dots, g_{N_{c_2},c_2}),$$

$$G'_i = G_i, \quad 1 \leq i \leq C, i \neq c_1, c_2.$$

**Table 5.1** Number of events for many-chromosome rearrangements.

Chromosomal rearrangement	No. of rearrangements
Inversion	$\sum_{i=1}^C \frac{N_i(N_i + 1)}{2}$
Shift	$\sum_{i=1}^C \frac{(N_i - 1)N_i(N_i + 1)}{6}$
Inverted shift	$\sum_{i=1}^C \frac{(N_i - 1)N_i(N_i + 1)}{6}$
Tandem duplication	$\sum_{i=1}^C \frac{N_i(N_i + 1)}{2}^\dagger$
Deficiency	$\sum_{i=1}^C \frac{N_i(N_i + 1)}{2}$
Inter-chromosomal translocation	$\sum_{i=1}^C \frac{(N + C - N_i - 1)N_i(N_i + 1)}{2}$
Inverted inter-chromosomal translocation	$\sum_{i=1}^C \frac{(N + C - N_i - 1)N_i(N_i + 1)}{2}$
Reciprocal translocation	$\sum_{i=1}^C \frac{(N - C + N_i + 1)(N_i - 1)}{2}$
Centric fusion	$C(C - 1)^\ddagger$
Dissociation	$4C^\ddagger$
Autopolyploidy	1
Allopolyploidy	1

<sup>†</sup>However, we may wish to limit the size of this event.

<sup>‡</sup>Due to the different potential orientations of the new gene orders.

An inverted inter-chromosomal translocation is similar to an inter-chromosomal translocation, but with the segment inverting before it is inserted into its new location. We define the gene orders following an inverted inter-chromosomal translocation  $INVINS(c_1, b_1, b_2; c_2, b_3)$  between breakpoints  $b_1$  and  $b_2$  on chromosome  $c_1$  and breakpoint  $b_3$  on chromosome  $c_2$ , where  $0 \leq b_1 < b_2 \leq N_{c_1}$ ,  $0 \leq b_3 \leq N_{c_2}$ ,  $1 \leq c_1, c_2 \leq C$  and  $c_1 \neq c_2$ , to be:

$$G'_{c_1} = (g_{1,c_1}, \dots, g_{b_1,c_1}, g_{b_2+1,c_1}, \dots, g_{N_{c_1},c_1}),$$

$$G'_{c_2} = (g_{1,c_2}, \dots, g_{b_3,c_2}, -g_{b_2,c_1}, \dots, -g_{b_1+1,c_1}, g_{b_3+1,c_2}, \dots, g_{N_{c_2},c_2}),$$

$$G'_i = G_i, \quad 1 \leq i \leq C, i \neq c_1, c_2.$$

A reciprocal translocation is a two-break mutation that occurs when parts of two chromosome arms swap with one another. We define the gene orders following a reciprocal translocation  $REC(c_1, b_1; c_2, b_2)$  between breakpoint  $b_1$  on chromosome  $c_1$  and breakpoint  $b_2$  on chromosome  $c_2$ , where  $0 \leq b_1 \leq N_{c_1}$ ,  $0 \leq b_2 \leq N_{c_2}$ ,  $1 \leq c_1, c_2 \leq C$  and  $c_1 \neq c_2$ , to be:

$$\begin{aligned} G'_{c_1} &= (g_{1,c_1}, \dots, g_{b_1,c_1}, g_{b_2+1,c_2}, \dots, g_{N_{c_2},c_2}), \\ G'_{c_2} &= (g_{1,c_2}, \dots, g_{b_2,c_2}, g_{b_1+1,c_1}, \dots, g_{N_{c_1},c_1}), \\ G'_i &= G_i, \quad 1 \leq i \leq C, i \neq c_1, c_2. \end{aligned}$$

A centric fusion or *Robertsonian translocation* is a two-break mutation where two distinct chromosomes join at their centromeres (with the very small remaining parts of the chromosomes above the centromeres being lost – these do not usually contain genes). We define the gene orders following a centric fusion  $FUS(c_1, b_1; c_2, b_2)$  between breakpoint  $b_1$  on chromosome  $c_1$  and breakpoint  $b_2$  on chromosome  $c_2$ , where  $0 \leq b_1 \leq N_{c_1}$ ,  $0 \leq b_2 \leq N_{c_2}$  (note that  $b_1$  and  $b_2$  must be the positions of the centromeres on chromosomes  $c_1$  and  $c_2$ , respectively),  $1 \leq c_1, c_2 \leq C$  and  $c_1 \neq c_2$ , to be:

$$\begin{aligned} G'_{c_1} &= (-g_{N_{c_1},c_1}, \dots, -g_{1,c_1}, g_{1,c_2}, \dots, g_{N_{c_2},c_2}) \\ \text{or } G'_{c_1} &= (-g_{N_{c_2},c_2}, \dots, -g_{1,c_2}, g_{1,c_1}, \dots, g_{N_{c_1},c_1}), \\ G'_{c_2} &= (), \\ G'_i &= G_i, \quad 1 \leq i \leq C, i \neq c_1, c_2, \\ C' &= C - 1. \end{aligned}$$

A dissociation (or *fission*) is a one-break mutation, where a single chromosome breaks at its centromere to produce two fully functional chromosomes. We define the gene orders following a dissociation  $DIS(c_1, b_1)$  at breakpoint  $b_1$  on chromosome  $c_1$ , where  $0 \leq b_1 \leq N_{c_1}$  (note that  $b_1$  must be the position of the centromere on chromosome  $c_1$ ) and  $1 \leq c_1 \leq C$ , to be:

$$\begin{aligned} G'_{c_1} &= (-g_{b_1,c_1}, \dots, g_{1,c_1}) \quad \text{or} \quad G'_{c_1} = (g_{1,c_1}, \dots, g_{b_1,c_1}), \\ G'_{C+1} &= (g_{b_1+1,c_1}, \dots, g_{N_{c_1},c_1}) \quad \text{or} \quad G'_{C+1} = (-g_{N_{c_1},c_1}, \dots, -g_{b_1+1,c_1}), \\ G'_i &= G_i, \quad 1 \leq i \leq C, i \neq c_1, \\ C' &= C + 1. \end{aligned}$$

The multi-chromosome non-conservative mutations are *polyploidy* events. A polyploidy is essentially a genome doubling. There are two kinds: autopolyploidy and allopolyploidy. An autopolyploidy event occurs when a genome gives rise to a new genome with two copies of each of its chromosomes. The new genome is an instant species that cannot interbreed with the species from which it derives. We define the gene orders following

an autopolyploidy  $AUP(G_i)$  of gene order  $G_i$  to be:

$$\begin{aligned} G'_i &= G_i, & 1 \leq i \leq C, \\ G'_{C+i} &= G_i, & 1 \leq i \leq C, \\ C' &= 2C. \end{aligned}$$

An allopolyploidy event occurs when a new genome receives the genomes of two distinct (but usually closely related) species. Again, the new genome belongs to a new instant species that cannot interbreed with either of the genomes from which it derives. We define the gene orders following an allopolyploidy  $ALP(G(1), G(2))$  of gene orders  $G_i(1)$  (for  $i = 1$  to  $C_1$ ) and gene orders  $G_j(2)$  (for  $j = 1$  to  $C_2$ ) to be:

$$\begin{aligned} G'_i &= G_i(1), & 1 \leq i \leq C_1, \\ G'_{C_1+j} &= G_j(2), & 1 \leq j \leq C_2, \\ C' &= C_1 + C_2. \end{aligned}$$

We may also wish in future to define other mutations, as we become aware of them. For example, O'Keefe and Eichler (2000) have recently discussed the possibility of a duplicative transposition event in human evolution, where duplications of large genomic segments are transferred to a new location in the genome.

### 5.3.3 Which Chromosomal Rearrangements Should We Consider?

We have seen that there are many types of chromosomal rearrangement and perhaps we should consider each type in an analysis of gene order and gene content data. However, most species have 'preferences' for certain types of chromosomal rearrangement. This may be because they have reproductive mechanisms that tolerate a particular mutation (e.g. the process of *pseudosynapsis*, which means that inversion heterozygotes give rise to balanced gametes) and therefore will not be as deleterious as other rearrangements or that they can be used for adaptive evolution (such as particular types of inversion in *Drosophila*).

In general, rearrangement heterozygotes will usually be less fit in the absence of such mechanisms, so that deleterious rearrangements are less likely to become fixed in a population and would be expected to be rare. Consequently, we often see the same type of rearrangement occurring again and again in a particular lineage. This was termed *karyotypic orthoselection* by White (1973). This means we may only wish to consider a small subset of chromosomal rearrangements in an analysis, if we have prior knowledge that this subset is important in the species being analysed. In general, the set of mutations that we will consider is known as the *evolutionary model*. If all of our considered rearrangements are conservative events then we have a conservative model. All other models are non-conservative.

## 5.4 MEASURING GENOME DIFFERENCE

Many of the problems outlined in Section 5.1 are tackled with reference to some measure of *genome difference* between a pair of genomes. Genome difference measures can be broadly grouped into two categories: *distance-based* and *path decomposition*. The first of these includes several fairly simple measures of the differences between gene orders and gene contents. The second category involves analysing likely sequences of mutations between gene orders with similar gene contents. However, as we will see, the two categories are not mutually exclusive and some distance measures are based upon the results of a path decomposition analysis. Here we will not present an exhaustive list of distance measures but rather give a flavour of the types of distance that have been proposed.

### 5.4.1 Distance Measures

When DNA sequences are being compared, the *edit distance* is defined as the smallest number of substitutions that transforms one sequence into another. This is fairly simple to calculate, and has been used successfully for many years. An analogous distance for gene content and gene order was proposed by Sankoff *et al.* (1992), one of the first distances for such data. This edit distance  $D_E$  is based on the *insertion/deletion distance*  $D_D$ , which is just the number of genes found in one genome that are not present in the other, and the *rearrangement distance*  $D_R$ , which is the minimum number of conservative evolutionary events required to convert one gene order into the other (ignoring differences in gene content), such that  $D_E = D_D + D_R$ . It was applied to a variety of plant, animal, and fungal mitochondrial genomes, and a least-squares algorithm was applied to the distance matrix to obtain a phylogenetic tree. Unfortunately, it was noted that the practical application of the algorithm was very slow, mainly due to the calculation of  $D_R$  (see Section 5.4.2).

In order to calculate distances more rapidly, Sankoff and Blanchette (1997; 1998) began using the *breakpoint distance*  $D_{BP}$ . This is the number of adjacent gene pairs in one gene order that are not adjacent (and in a similar orientation) in a second. At about the same time, Dicks (1999) began using the closely related *segment distance*  $D_{SEG}$ , the number of mutually conserved segments between two gene orders with identical content. Sankoff and Blanchette further developed the *induced breakpoint distance* to allow the comparison of very differently sized genomes. Here, genes that are not common to both species are removed before calculating the distance. This in turn led to the *normalized induced breakpoint distance* (Sankoff *et al.*, 2000a), where the induced breakpoint distance is normalized by dividing the distance by the number of the remaining shared genes. The normalized induced breakpoint distance was independently developed by Herniou *et al.* (2001).

Gene adjacency conservation was used as a simple measure by Keogh *et al.* (1998) to compare different fungal genomes. The neighbour pair distance  $D_{NP}$  is simply the proportion of adjacent gene pairs in one genome that are adjacent in the second. Cosner *et al.* (2000) also looked at neighbouring pairs but used them as binary characters, known as a *binary encoding*. Pairs that were adjacent (and of consistent orientation) were scored as 1, and those that were not as 0. This encoding, which could also be weighted if wished, could then be analysed using a standard maximum parsimony analysis. Again, this binary encoding was independently developed by Herniou *et al.* (2001) in their comparative

analysis of DNA sequence, gene order and gene content in inferring phylogenetic trees for a baculovirus data set.

The distances above are not natural to use when one genome contains two homologues of a gene in another genome, an *orthologue* (the original homologue) and a *paralogue* (a duplicate of the orthologue). Sankoff (1999) defined the *exemplar distance* for use in this situation. We must first determine which gene is the orthologue and which the paralogue, which is not always possible to do or, when done, not always correct. Sankoff chooses to delete the gene which leads to the smallest distance between the reduced genomes. For some types of distance this measure can be computationally expensive to find, as there may be many such genes to delete, but there are no plausible alternatives at present. The exemplar distance can be calculated based on any of the above distances (e.g. the exemplar breakpoint distance or the exemplar edit distance).

In addition to distances influenced largely by gene order, there are now several based on gene content. Gu (2000) developed a stochastic model of gene loss, using a death process, and showed how it could be used to develop a content-based distance between two genomes. Snel *et al.* (1999) proposed that the percentage of genes shared by two genomes could be used as a measure of similarity between them. Ferretti *et al.* (1996) proposed the *syntenic distance*, based on the content of each chromosome. Here, the minimum number of centric fusions, dissociations and translocations required to ensure that every chromosome within one genome has the same gene content as its homologue, ignoring gene order, is counted.

Some of the distance measures defined above are related simply for single chromosomes with identical gene content:

$$D_{BP} = D_{SEG} + 1 = N - (N \times D_{NP}) - 1.$$

However, this relationship becomes more complicated when we have differing content and duplicated genes. A comprehensive investigation into the behaviour and performance of these measures under different conditions has yet to be carried out. Furthermore, it would be preferable to consolidate gene content and gene order into a single distance.

#### 5.4.2 Path Decomposition

The sequence of chromosomal rearrangements that mutates the gene order  $G_i$  to gene order  $G_j$  is known as an *evolutionary path*  $\rho_{i,j}$ . Finding such a path or paths between two genomes is known as path decomposition analysis. The most widely used method of performing a path decomposition, to find paths separating a *query genome* and a *target genome*, has been used by Blanchette *et al.* (1996) and by Dicks (1999), amongst others. Here we mutate our query genome in all possible ways, according to our evolutionary model. We are left with  $E$  new gene orders (where  $E$  is the number of potential events in our evolutionary model). We then mutate the  $E$  gene orders, each according to our  $E$  mutations, to give us  $E^2$  new gene orders. We repeat this process, the result being a branching search tree. At each node of the search tree we check to see if we have a genome identical to our target genome. If this is the case, we have achieved a successful path.

For the simple case of a conservatively evolving single chromosome of size  $N$  and for a search tree of depth  $l$ , we find that the number of events  $E$  is

$$E = \binom{N+1}{2} + 2\binom{N+1}{3} = N(N+1)(2N+1)/6,$$

and the number of nodes in the search tree is

$$\sum_{k=1}^l E^k = \frac{E(E^l - 1)}{E - 1}.$$

Obviously, the number of nodes quickly becomes very large, even for small  $N$ . This is greatly pronounced in multi-chromosome genomes, and indeed  $E$  is not a constant in this case and so the search tree can become quite complex. In practice, the tree must be pruned according to some heuristic. The authors cited above differ slightly in their pruning strategies. Furthermore, they differ more markedly in the goals of their search. Blanchette *et al.* look for the shortest path(s) separating the query genome from the target genome. Blanchette *et al.* prune the search tree with  $D_{BP}$  and use a depth-first search with limited look-ahead, and branch and bound for a rapid search. Dicks uses the related  $D_{SEG}$  distance measure to prune the tree, again with a depth-first search, but here the search does not stop once the shortest path is achieved. The computation is stopped once the tree has been completely pruned or once a conservative upper bound on tree depth is achieved, whichever occurs first. The result is a *set of paths* separating the two genomes.

Once we have carried out such an analysis and found a set of paths  $R$  between  $G_i$  and  $G_j$ , we can use the results to calculate a transition probability, under some probabilistic model  $M$  (for example, including parameters for time  $t$  and rate of mutation  $\lambda$ ), between the two genomes. We can define this (setting the constant of proportionality to one) as

$$L(\lambda, t|G_i, G_j) = \sum_R P(\rho_{i,j}|\lambda, t),$$

which is often approximated by

$$L(\lambda, t|G_i, G_j) = \max_R P(\rho_{i,j}|\lambda, t).$$

The former approach uses all paths to calculate the transition likelihood. Note that the real set of paths separating two gene orders is of infinite size, and that we approximate the real transition likelihood by summing over  $R$ . A combined likelihood approach was shown by Thorne *et al.* (1991; 1992) to be less biased than a single maximum likelihood in the analysis of DNA sequence alignment and, although not formally tested for chromosomal evolution, it is likely to be the case here. The latter definition uses the maximum likelihood path  $\hat{\rho}_{i,j}$ , which is often used for computational convenience. Under any sensible model of evolution, where  $\ell(\rho)$  is the length of a path  $\rho$ ,

$$\lim_{\ell(\rho) \rightarrow \infty} P(\rho) = 0.$$

Ideally, we would like to use this be able to choose just those paths that make a significant contribution to the combined transition likelihood. This would be equivalent to using path

likelihood, under a particular probabilistic model, as a pruning statistic. Unfortunately this is not computationally feasible at present, but it does warrant further investigation.

The path or paths resulting from a path decomposition analysis can be used directly to answer evolutionary questions. Furthermore, we can use a path decomposition analysis to obtain a distance measure between two genomes. Firstly, we note that the number of events in the shortest path is simply the rearrangement distance  $D_R$ . Secondly, we can calculate a probabilistic distance measure, based on the path decomposition. For example, for each genome pair, where our data are evolving under some stochastic model  $M$ , we can find a maximum likelihood estimate for time,  $\hat{t}$ , as the distance between our two genomes. Durbin *et al.* (1998) showed that maximum likelihood estimates of distance are additive, lending themselves well to a subsequent phylogenetic analysis. It should be noted that the shortest path will not always be the maximum likelihood path, although Savva (2001) has shown that, in general, just a few short paths make a very large contribution to the combined transition likelihood.

## 5.5 STATISTICAL MODELS OF CHROMOSOMAL EVOLUTION

Markov chains and, more recently, hidden Markov models have been used widely to model the evolution of DNA and amino acid sequences. For DNA sequences, where each nucleotide site is considered independently, the Markov property lends itself well to modelling the transitions between the four possible states A, C, G and T in time  $t$ . Even when considering pairs of adjacent bases, as has been done for modelling RNA sequences, we need only consider 16 possible states. Amino acid evolution is a little more complicated, with 20 states to consider. We can see that gene orders also move between states and it would be sensible to consider an analogous model for their evolution. However, a gene order is a permutation, where adjacent genes are not independent. Imagine a small chromosome consisting of just 10 genes. If we wished to use a Markov chain model, we would need to consider  $10! \times 2^{10}$  or a little over 3.7 billion states. Obviously, this is computationally infeasible, even for such a small data set. Consequently, we cannot use Markov chains and hidden Markov models to directly model gene order state space. However, they could be used to model properties of gene orders, as opposed to the gene orders themselves (Dicks, work in progress).

### 5.5.1 Modelling Paths

Before looking at a simple model of chromosomal evolution, we should consider the issue of *reversibility*. This is important if we do not have prior knowledge about the direction of evolution within our data set and, for instance, we wish to estimate a phylogenetic tree (see Section 5.6). If we choose not to distinguish between members of a rearrangement class then this issue is quite simple. However, if we choose to develop subclasses of rearrangement events (see Section 5.9) then we must take care with this issue. For example, we know there are  $\binom{N+1}{2} = N(N+1)/2$  possible inversion events for a chromosome with  $N$  genes, and most models impose upon them that they are equally likely to occur, irrespective of the locations of the two breaks. Under such a model, it is simple to see that an inversion is its own inverse event and that the likelihood of a path going forwards in time would be equal to one going backwards. Indeed,

for the conservative rearrangements inversion, shift, inverted shift, inter-chromosomal translocation, inverted inter-chromosomal translocation and reciprocal translocation, each type of event is its own inverse type.

The remaining conservative events, centric fusion and dissociation, are each other's inverse event type. Consequently, as long as each evolutionary model that includes centric fusion also includes dissociation, and *vice versa*, and we do not distinguish between individual events, then reversibility of the model will be guaranteed. The non-conservative case is a little more complicated. A deficiency is the inverse of a tandem duplication, but we need several deficiencies to reverse a polyploidy event. This could be resolved by developing a multi-chromosome loss event or by adjusting the rates of our mutations, but are these solutions based on a biological reality? We must take care when modelling our events that we stay as close to real biological events as we can. Consequently, we must define our deficiency events carefully to ensure reversibility. An additional problem stems from our searching methods in a path decomposition analysis. We must ensure that pruning does not prune the inverse of a path that would not be pruned if we swapped our query and target genomes. If this were the case, then reversibility would break down.

Let us develop a simple model of chromosomal evolution and show how it could be used in both a traditional maximum likelihood analysis and a Bayesian analysis. For simplicity, the model will be applied to the single-chromosome conservative case, but it could be extended for more complex cases with little difficulty. Waiting times between chromosomal rearrangements are exponentially distributed with rates  $\lambda_I$  for inversions and  $\lambda_S$  for shifts and inverted shifts. So the numbers of each event occurring in a time  $t$  have independent Poisson distributions with parameters  $\lambda_I t$  and  $\lambda_S t$  respectively, and the total number of events has a  $Po(\lambda t)$  distribution with  $\lambda = \lambda_I + \lambda_S$ . So, for any path,  $\rho$ , separating gene orders  $G_i$  and  $G_j$ , the probability that  $\ell(\rho)$  events occurred in  $t$  is

$$P_t(\ell(\rho)) = \frac{e^{-\lambda t} (\lambda t)^{\ell(\rho)}}{\ell(\rho)!}.$$

We also require that these events are the particular inversion and shift events found within the path, and we assume that each inversion is equally likely (similarly for shifts). Since each event occurs independently of the others we find that, for  $i(\rho)$  inversions and  $s(\rho)$  shifts and inverted shifts along  $\rho$ :

$$P_t(\rho) = \frac{e^{-\lambda t} (\lambda t)^{\ell(\rho)}}{\ell(\rho)!} \left( \frac{\lambda_I}{\lambda_I + \lambda_S} \frac{2}{N(N+1)} \right)^{i(\rho)} \left( \frac{\lambda_S}{\lambda_I + \lambda_S} \frac{3}{N(N+1)(N-1)} \right)^{s(\rho)}.$$

Notice that this probability factorizes into the Poisson probability depending on  $t$ , and the rest depending only on  $\rho$ . We write this second factor as  $w(\rho)$ , the *weight* of the path. So

$$P(G_i \rightarrow G_j | \lambda, t) = \sum_R \left( \frac{e^{-\lambda t} (\lambda t)^{\ell(\rho)}}{\ell(\rho)!} w(\rho) \right).$$

We can use this model in, for example, a Bayesian analysis to find a posterior distribution for  $T$ , the time of a path. If we use the improper  $\Gamma(0, 0)$  distribution as a prior

for our random variable  $T$ , then Bayes' theorem gives us the posterior density for  $T$ :

$$f(t) \propto t^{-1} \sum_R \left( \frac{e^{-\lambda t} (\lambda t)^{\ell(\rho)}}{\ell(\rho)!} w(\rho) \right) = \sum_R \left( \frac{\lambda^{\ell(\rho)}}{\ell(\rho)!} e^{-\lambda t} t^{\ell(\rho)-1} w(\rho) \right),$$

which is a mixture of gamma distributions. This can be used to find the expected value of  $T$ :

$$E(T) = \frac{1}{\lambda} \frac{\sum_R \ell(\rho) w(\rho) / \ell(\rho)}{\sum_R w(\rho) / \ell(\rho)} = \frac{1}{\lambda} \frac{\sum_R w(\rho)}{\sum_R w(\rho) / \ell(\rho)},$$

which is a weighted average of the lengths of the paths. We could similarly find the variance or probability intervals for  $T$ . For full details of this model, see Savva (2001).

There is a small number of chromosomal models in the literature, some of which can be used to model paths. Sankoff and Ferretti (1996) describe a model of random reciprocal translocation for a fixed number of chromosomes or chromosome arms and use it to find the equilibrium distance of chromosome lengths. Dicks (1999) proposed the iterative segment model, a description of the number of segments in an evolutionary path, and the exponential failure model (similar to that shown above), which uses a Poisson process to model the number of mutations in a path. Gu (2000) uses a death process to describe gene loss and shows how it can be used to infer a gene content phylogeny using both distance-based and maximum likelihood approaches. Recently, Larget *et al.* (2002) devised a model for inversion within conservative single chromosomes that could be used in a Bayesian phylogenetic analysis. The model defined a gamma distribution for branch lengths, a Poisson process for the number of inversion events on tree branches, and uniform distributions for the times between events on branches and tree topologies. At present, none of these models is used routinely in comparative genomic studies. Now that we are beginning to have a choice of model of chromosomal evolution, we need to test whether these and future models are adequate descriptions of chromosomal evolution for a range of data sets, as has been shown for models of DNA sequence evolution by Yang *et al.* (1994).

### 5.5.2 Modelling Distances

The maximum likelihood distance approach has been investigated by a few researchers. Dicks (1999; work in progress) used the iterative segment model to find the likelihood of a path (for both the single- and multi-chromosomes cases, undergoing conservative evolution), given an observed value of  $D_{\text{SEG}}$ . Caprara and Lancia (2000) showed some interesting results, such as the expected number of breakpoints in a random permutation of  $N$  genes,

$$E[D_{\text{BP}}] = N - 1,$$

and the expected number of breakpoints in a random permutation of  $N$  genes after a random path of  $k$  inversions,

$$E[N(k)] = (N - 1) \left( 1 - \left( \frac{N - 3}{N - 1} \right)^k \right).$$

They used these results to calculate the conditional probability of the number of inversions given an observed value of  $D_{BP}$ . Savva (work in progress) has also investigated the joint conditional probability distribution of an observed value of  $D_{SEG}$  and a calculated shortest path for a given path length.

### 5.5.3 Limitations of Gene Order Analysis

Path probabilities tend to a uniform ergodic distribution. This means that, once a particular evolutionary time has passed or a number of events has occurred, the probability of one path is equal to that of any other. Evolutionary analyses under such conditions would be meaningless. However, this may not be a problem as long as the number of mutation events does not lead us close to this point. Dicks has used the iterative segment model to estimate when stationarity is reached (to a given level of accuracy), for a single chromosome evolving conservatively, and found that it is linear with the number of genes in the chromosome. Savva found that, for large  $t$ , the gene content overtakes gene order in providing information for evolutionary analyses. This must be investigated more thoroughly, particularly as it is probable that more work will be carried out on nuclear genomes in the near future. Here, the number of mutations will be larger than that seen for mitochondrial genomes on which most research has focused to date (although the number of genes will also be much larger).

## 5.6 PHYLOGENETIC ANALYSIS

It would be pleasing to think that chromosomal phylogenetic trees, which represent the evolution of the genome as a whole, more closely mirror the evolution of species than the evolution of a single gene sequence. While this is an attractive notion, we do not have conclusive evidence to confirm it. However, comparing different sorts of trees may throw up interesting biological questions. For example, genes that evolve very differently from chromosomes may be of particular interest. A recent comparison of gene sequence, gene order and gene content data sets for use in the phylogenetic analysis of baculoviruses (Herniou *et al.*, 2001) showed how these different approaches could be used to lend weight to taxonomic hypotheses.

Chromosomal banding patterns gave rise to what was probably the first comparative genome analysis by Dobzhansky and Sturtevant (1938), where the banding patterns of *Drosophila* were used to derive an evolutionary tree. This analysis was carried out by eye, which was made possible by the relative simplicity of the data. Today, we wish to carry out similar analyses, but for more complex data sets and in a systematic manner.

A phylogenetic tree can be calculated easily using one of the distance measures described in Section 5.4. A distance matrix is constructed, where each entry in the matrix is the distance between a pair of genomes. This matrix is used as input to a tree-building algorithm. The neighbour-joining algorithm (Saitou and Nei, 1987) is perhaps the best known of these but others such as the FITCH (Fitch and Margoliash, 1967) and KITSCH algorithms within the PHYLIP phylogenetic package (Felsenstein 1993) are also applicable. Each of these algorithms has its own advantages and disadvantages, as has been described frequently elsewhere.

We have also seen in Sections 5.4 and 5.5 that we can calculate transition probabilities between two gene orders and gene contents. Therefore, it should be possible to calculate a maximum likelihood tree in a similar way to that described by Felsenstein (1981) for DNA sequences. Here, for a hypothesis  $H$  of how the bases of the sequences evolved (the model of evolution), a tree structure  $\tau$ , internal tree nodes  $I$ , tree branches of length  $t$  and observed sequence data  $D$  (of bases  $d_j (j = 1, \dots, N)$ ), where the sequence consists of  $N$  bases, the likelihood of the hypothesis, given the observed data, is

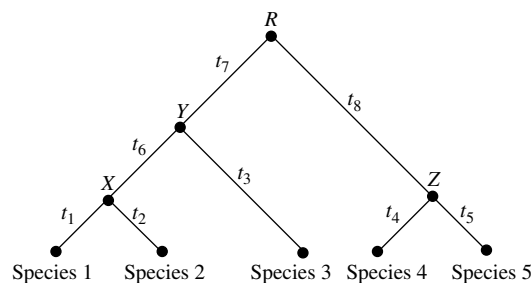
$$L(H|D) = L(\tau, t, I|D) = \prod_{j=1}^N P(d_j|\tau, t, I).$$

Felsenstein shows how his method treats the states at the internal tree nodes as nuisance parameters and integrates over them. For example, consider the rooted phylogenetic tree in Figure 5.2. The tree shows a root,  $R$ , three internal nodes  $X$ ,  $Y$  and  $Z$  and branch lengths  $t_1, \dots, t_8$ . Let  $r, x, y$  and  $z$  represent the states of the nodes  $R, X, Y$  and  $Z$ , respectively. The likelihood of this tree (for any type of data, including DNA sequences, gene orders etc.), would be:

$$L = \sum_r \pi_r \left( \sum_z P_{t_8}(r \rightarrow z) P_{t_4}(z \rightarrow 4) P_{t_5}(z \rightarrow 5) \right) \\ \times \left( \sum_y P_{t_7}(r \rightarrow y) P_{t_3}(y \rightarrow 3) \left( \sum_x P_{t_6}(y \rightarrow x) P_{t_1}(x \rightarrow 1) P_{t_2}(x \rightarrow 2) \right) \right),$$

where  $\pi_r$  represents the probability of the root being in state  $r$  (for DNA sequences, observed base frequencies are used to calculate these prior probabilities). For DNA sequences, this likelihood would be found for each base site where  $R$  could take one of four states,  $A, C, G$  and  $T$ . The  $N$  likelihoods would be multiplied together to form the overall tree likelihood. Felsenstein also shows how to calculate the likelihood of an unrooted tree, using the ‘pulley principle’ for a reversible model of evolution.

However, there are major differences between DNA sequences and gene orders and, to a lesser extent, gene contents. For the latter two, we only need calculate a single likelihood value as our genes are not independent entities as are our nucleotide sites. However, for a chromosomal tree, the root and the internal nodes can take any one of billions of potential states, and in order to calculate the likelihood shown above we would have



**Figure 5.2** An example of a phylogenetic tree

to sum over these states. Obviously, this is not computationally feasible. Dicks (2000) shows a computationally plausible approximation to the maximum likelihood tree for gene order data, using a small set of states at internal nodes, where those states are thought to make a large overall contribution to the tree likelihood. Markov chain Monte Carlo methods have been used to search through tree space for the maximum likelihood tree (e.g. Larget and Simon, 1999) in the case of DNA sequences. Larget *et al.* (2002) also showed recently how Markov chain Monte Carlo methods could be applied to gene order problems. However, these approaches to estimating maximum likelihood trees for gene order data are still relatively new and there is much work to be done before they are used more widely. Consequently, almost all gene order and gene content trees are currently found via distance methods.

## 5.7 COMPUTATIONAL COMPARATIVE GENOME ANALYSIS

In this chapter we have naturally focused on the use of statistics to compare genomes. However, further to the research outlined above, there is a wealth of computational research in this area, including polynomial-time algorithms to find shortest paths between two gene orders using inversion only (e.g. Hannenhalli and Pevzner, 1995; Kaplan *et al.*, 1997; Bergeron, 2001); algorithms to solve the multiple genome rearrangement problem (Bourque and Pevzner, 2002) (finding the phylogeny, including internal nodes, that minimizes the number of mutations within a whole phylogenetic tree); and Steiner tree algorithms (finding the tree of minimum cost that includes all our gene orders – known as the *special* Steiner nodes) to find phylogenetic trees of gene orders (Koutsikou *et al.*, personal communication). At present, it is not possible to, for example, place confidence intervals on the estimates obtained from these analyses. However, a thorough further investigation to place them in a statistical framework would be highly useful, as these methods are, in general, capable of faster computation than the statistical methods described in this chapter.

## 5.8 COMPARATIVE GENOMICS SOFTWARE

The quantity of software to carry out analyses such as those described above has increased significantly in the past couple of years. We have seen that we can carry out pairwise analyses on distance matrices with well-established packages such as PHYLIP, using algorithms such as NEIGHBOR, FITCH and KITSCH.

The first program to focus specifically on gene order and gene content analysis was DERANGE, developed in the early 1990s. Version 2 of this software calculates weighted edit distances (Blanchette; <http://www.cs.washington.edu/homes/blanchem/software.html>) between pairs of gene orders, for inversions and shifts. This software was followed by BPAanalysis, also from the Sankoff group (Blanchette; <http://www.cs.washington.edu/homes/blanchem/software.html>), to estimate phylogenies and ancestral states from gene orders, optimizing on breakpoint distance. A recent package, GRAPPA (Moret *et al.*; <http://www.cs.unm.edu/~moret/GRAPPA/>), claims to calculate trees based

on breakpoint distance more rapidly than BPAAnalysis, in addition to estimating trees using the inversion distance.

GOTREE (Bryant and Johnson; <http://www.mcb.mcgill.ca/~bryant/GoTree/>) calculates breakpoint distances for gene orders with both equivalent and differing gene contents, and estimates phylogenetic trees using neighbour-joining. MGR finds a phylogenetic tree (Bourque; <http://www.cs.ucsd.edu/groups/bioinformatics/MGR/>) that minimizes the sum of chromosomal rearrangements over all edges of the tree. It can also be used to infer ancestral gene order states. It is based on the GRIMM (Tesler; <http://www.cs.ucsd.edu/groups/bioinformatics/GRIMM/>) software for calculating shortest distances between pairs of genomes.

CHROMTREE calculates pairwise distances, such as the segment distance, for input into pairwise tree generation programs. It also approximates maximum likelihood trees (Dicks; <http://jic-bioinfo.bbsrc.ac.uk/bioinformatics-research/software/CHROMTREE/>), as described in Section 5.6, under various models of chromosomal evolution. SHOT takes genome sequences as its data (Korbel *et al.*; <http://www.bork.embl-heidelberg.de/~korbel/SHOT/>) rather than gene orders. It first finds homologous genes and subsequently estimates trees using various pairwise distances. It can also perform tree bootstrapping using the jackknife.

## 5.9 CONCLUSIONS AND FUTURE RESEARCH

We have seen that there are many promising strands of research in comparative genome analysis, which will play an important role in making sense of the enormous quantities of data to be output in the coming decade. In particular, when we hear of new genome sequencing projects, we often hear that one of reasons for choosing the organism concerned was its close relationship to an already sequenced genome, so that comparative genome analysis could be carried out. However, we should answer several fundamental questions concerning such analyses in order to be able to understand what the results mean. Here we will touch upon a few of these questions.

The most important question is to ask what the analysis of gene order and gene content means. Are we sure that chromosomal evolution mirrors, or even drives, the evolution of species? Are we using the right data sets? At present, many analyses focus on the very large or *gross* rearrangements and ignore the more frequent smaller rearrangements, sometimes because the resolution of the data will not allow this at present. However, the small events are more likely to be random than the very large events, and the latter must be under enormous selective pressure as they can separate functional clusters of genes. Should we be looking at the small events alone or should our models distinguish between small and large events?

Are our models of evolution adequate? Currently, most models do not make use of the centromeres or telomeres, even though these play such an important role in chromosomal mutation. Should we add concepts such as karyotypic orthoselection? More importantly, can we verify particular models for particular data sets? Model verification is extremely important and yet is rarely carried out. We must take the lead from those researchers who have attempted to do this for DNA sequence evolution. This will also mean that we should develop further the maximum likelihood methodologies for gene order and gene content data. At present, these represent a fairly small proportion of work in this area.

Is there evidence for a chromosomal clock? At present, it does not appear likely that a chromosomal clock is keeping time behind the chromosomal mutations, at least for the very large ‘visible’ ones. There are many cases, in both the animal and plant kingdoms, where we see an extraordinary variation in the amount of chromosomal rearrangement in similar evolutionary time. This holds for closely related species as well as more divergent ones. Are phylogenetic trees adequate for our data sets? Certainly for species where allopolyploidy has taken place, we may need to consider estimating a phylogenetic network, as has been examined for DNA sequences by Strimmer and Moulton (2000). Can we place confidence intervals on our estimates? Although bootstrapping is routinely carried out on DNA sequence trees (Felsenstein, 1985), an analogous method for chromosomal trees is not obvious, due to the combinatorial nature of the data. Savva (work in progress) proposes parametric bootstrapping to find confidence intervals for phylogenetic trees, although this has not yet been implemented, and we mentioned in Section 5.8 that Korbel *et al.* have used the jackknife in the SHOT software.

Chromosomal rearrangements are highly correlated with the underlying DNA sequence, in particular the presence of transposable elements. Should we examine chromosomal evolution with reference to DNA? This will obviously make analyses even more complicated than at present, but would the results make this extra effort worthwhile? For example, from a preliminary analysis of our data, we would know that some breakpoints were much more likely than others and we could establish a statistical model with breakpoint hotspots. However, we would need a whole genome sequence to be able to carry out such an analysis.

Fortunately, we see a burgeoning interest in the area of comparative genomics. We note a growing number of ideas from biologists, statisticians and computer scientists. This is undoubtedly influenced by the ever increasing number of genome sequencing projects and also by a welcome emphasis on interdisciplinary research. Consequently, we can look forward to seeing the ideas and methods within comparative genomics mature over the coming decade.

### Acknowledgments

Thanks must go to the Biotechnology and Biological Sciences Research Council and the John Innes Foundation for supporting the research of JD and GS, respectively. Thank you to our many colleagues whose insight has contributed to this work over the past decade.

### REFERENCES

- Bergeron, A. (2001). In *Combinational Pattern Matching: 12th Annual Symposium*, Lecture Notes in Computer Science 2089, A. Amir and G.M. Landau, eds. Springer-Verlag, Berlin, pp. 106–117.
- Blanchette, M., Kunisawa, T. and Sankoff, D. (1996). Parametric genome rearrangement. *Gene-Combis* **172**, 11–17.
- Bourque, G. and Pevzner, P.A. (2002). Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research* **12**, 26–36.
- Caprara, A. and Lancia, G. (2000). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 171–184.
- Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.-S., Warnow, T. and Wyman, S. (2000). In *ISMB-2000: Proceedings, Eighth International Conference on Intelligent*

- Systems for Molecular Biology*, P. Bourne, M. Gribskov R. Altman *et al.*, eds. AAAI Press, Menlo Park, CA, pp. 104–115.
- Dicks, J.L. (1999). *Comparative mapping and phylogeny*. DPhil thesis, University of Oxford.
- Dicks, J. (2000). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 333–342.
- Dobzhansky, T. and Sturtevant, A.H. (1938). Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics* **23**, 28–64.
- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press, Cambridge.
- Ehrlich, J., Sankoff, D. and Nadeau, J.H. (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics* **147**, 289–296.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.
- Felsenstein, J. (1985). Confidence limits of phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
- Felsenstein, J. (1993). PHYLIP Version 3.5, <http://evolution.genetics.washington.edu/phylip.html>.
- Ferretti, V., Nadeau, J.H. and Sankoff, D. (1996). In *Combinatorial Pattern Matching: 7th Annual Symposium*, Lecture Notes in Computer Science 1075, D. Hirschberg and G. Meyers, eds. Springer-Verlag, Berlin, pp. 159–167.
- Fitch, W.M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279–284.
- Gu, X. (2000). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 515–524.
- Hannenhalli, S. and Pevzner, P. (1995). In *Proceedings of the Twenty-Seventh Annual ACM–SIAM Symposium in the Theory of Computing*. ACM Press, New York, pp. 178–179.
- Herniou, E.A., Luque, T., Chen, X., Vlaskovits, J.M., Winstanley, D., Cory, J.S. and O’Reilly, D.R. (2001). Use of whole genome sequence data to infer baculovirus phylogeny. *Journal of Virology* **75**, 8117–8126.
- Kaplan, H., Shamir, R. and Tarjan, R. (1997). In *Proceedings of the Eighth Annual ACM–SIAM Symposium on Discrete Algorithms*. ACM Press, New York, pp. 344–351.
- Keogh, R.S., Seoighe, C. and Wolfe, K.H. (1998). Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi. *Yeast* **14**, 443–457.
- Larget, B. and Simon, D. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* **16**, 750–759.
- Larget, B., Simon, D.L. and Kadane, J.B. (2002). Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *Journal of the Royal Statistical Society B* **64**, 681–693.
- Nadeau, J.H. and Taylor, B.A. (1984). Lengths of chromosomal segments conserved since divergence of mouse and man. *Proceedings of the National Academy of Sciences (USA)* **81**, 814–818.
- O’Keefe, C. and Eichler, E. (2000). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 29–46.
- Saitou, N. and Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.
- Sankoff, D. (1999). Genome rearrangements with gene families. *Bioinformatics* **15**, 909–917.
- Sankoff, D. and Blanchette, M. (1997). In *Computing and Combinatorics: Third Annual International Conference, COCOON ’97*, Lecture Notes in Computer Science 1276, T. Jiang and D.T. Lee, eds. Springer-Verlag, Berlin, pp. 251–263.
- Sankoff, D. and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology* **5**, 555–570.
- Sankoff, D. and Ferretti, V. (1996). Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Research* **6**, 1–9.
- Sankoff, D. and Nadeau, J.H. (eds) (2000). *Comparative Genomics*. Kluwer Academic, Dordrecht.

- Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F. and Cedergren, R. (1992). Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences (USA)* **89**, 6575–6579.
- Sankoff, D., Deneault, M., Bryant, D., Lemieux, C. and Turmel, M. (2000a). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 89–98
- Sankoff, D., Parent, M.-N. and Byant, D. (2000b). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 299–306.
- Savva, G. (2001). *Estimating transition probabilities using a model of chromosomal evolution*. MSc dissertation, University College London.
- Schoen, D.J. (2000). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 307–320.
- Snel, B., Bork, P. and Huynen, M.A. (1999). Genome phylogeny based on gene content. *Nature Genetics* **21**, 108–110.
- Strimmer, K. and Moulton, V. (2000). Likelihood analysis of phylogenetic networks using directed graphical models. *Molecular Biology and Evolution* **17**, 875–881.
- Thorne, J.L., Kishino, H. and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33**, 114–124.
- Thorne, J.L., Kishino, H. and Felsenstein, J. (1992). Inching toward reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34**, 3–16.
- Waddington, D. (2000). In *Comparative Genomics*, D. Sankoff and J.H. Nadeau, eds. Kluwer Academic, Dordrecht, pp. 321–332.
- Waddington, D., Springbett, A.J. and Burt, D. (2000). A chromosome based model for estimating the number of conserved segments between pairs of species from comparative genetic maps. *Genetics* **154**, 323–332.
- White, M.J.D. (1973). *The Chromosomes*, 6th edition. Chapman & Hall, London.
- Yang, Z., Goldman, N. and Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution* **11**, 316–324.