A Small Trip in the Untranquil* World of Genomes

A survey on the detection and analysis of genome rearrangement breakpoints

Claire Lemaitre [1,2,◇] and Marie-France Sagot[1,2]

January 1, 2007

[1] Équipe BAOBAB, Laboratoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS; Univ. Lyon 1, 43 bd du 11 nov 1918, 69622, Villeurbanne Cedex, France.
[2] Projet Helix, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France
[◇] Corresponding author (clemaitr@biomserv.univ-lyon1.fr)

Abstract

Genomes are dynamic molecules that are constantly undergoing mutations and rearrangements. The latter are large scale changes in a genome organisation that participate in the evolutionary and speciation process, but may also be involved in inherited diseases and in cancer. They have since long been studied by the biologists whereas computational biologists have since more recently only been attracted to the topic.

One of the (exciting) objectives for studying rearrangements is to understand the underlying molecular mechanisms of evolution. One possible line of investigation is to analyse, at the sequence level, the regions which have undergone a rearrangement, assuming we are able to very precisely locate them.

This paper presents a survey of the different methods that have been developed to identify such regions, in particular the approaches that are based on the alignment of genomes. The main purpose of the paper is then to investigate what is currently known about the characteristics of the regions where a rearrangement took place, and about the mechanism(s) having led to such large scale changes.

keywords: Genome dynamics, rearrangement, breakpoint region, whole genome alignment, conserved segment, synteny block

1 Introduction

From a relatively marginal topic when others, like sequence alignment for instance, were in full bloom in the early years of computational biology, genome dynamics has evolved into an increasingly more active area of research. The area has grown also in sophistication although the models used remain in general biologically unrealistic. Indeed, this is an area where the gap between what has been done by the computational biologists, and what has long been known or is believed to be true

*Then let us clear away the choaking thorns/ From round its gentle stem; let the young fawns,/ Yeaned in after times, when we are flown,/ Find a fresh sward beneath it, overgrown/ With simple flowers: let there nothing be/ More boisterous than a lover's bended knee;/ Nought more ungentle than the placid look/ Of one who leans upon a closed book;/ Nought more **untranquil** than the grassy slopes/ Between two hills. All hail delightful hopes! – from *The Poetical Works of John Keats*.

by the biologists is perhaps the greatest in the field. Algorithmicists in particular, and among them those coming from a combinatorial background, have loved the problem in its initial formulations because of its close relation to concepts long familiar to them, such as permutations, and because of the simplicity of the questions one can ask. This was thus a topic where it seemed possible to make important contributions without having to get too deeply into the underlying biological complexity. This is, of course, not true, at least not anymore as soon as one starts wanting to "interpret" the results obtained or to use them to further our knowledge on, for instance, evolution.

This issue is one of the causes of some very recent polemics concerning genome dynamics. Part of the polemics (for instance, surrounding the issue of the existence or not of "hotspots" – regions along a genome that are more susceptible to be the loci of rearrangements [**?**] [**?**] [**?**] [**?**] [**?**]) have involved different groups of computational biologists. Others, such as examplified in the March issue of *Genome Research* [**?**] [**?**], have involved computational biologists with biologists working with data (coming from cytogenetics[1]) that pre-existed the sequencing of whole genomes.

Our purpose with this survey is not to participate ourselves in those polemics, nor to explore all the aspects behind genome dynamics. Indeed, a whole book would not be enough for this. We shall concentrate instead on two questions. The first is detecting the breakpoints, that is the exact points along a genome where a rearrangement has taken place, in the organism under study or in the homologous genome locus of another organism. The second question concerns the analysis of the regions around breakpoints.

In fact, the first question, detecting the breakpoints, which seems very simple, is a hard one, and, to the best of our knowledge, has never been addressed in this very precise way. Many approximations have been made in the sense that all methods that detect genome segments conserved among different organisms are trying to delimit a more or less wide region around possible breakpoints. Part of this paper will be a survey of such methods, and of methods developed for another purpose but that could be used to identify conserved segments, and thus, their duals, that is regions that at some point were "broken". It is ironic, and satisfying, that this will take us back to the very beginnings of computational biology: sequence alignments! The scales though are not at all the same anymore.

Why an interest for precisely detecting breakpoints? One main motivation does lead us back to some of the polemics we alluded to above: finely analysing the regions around breakpoints could give us some clues on the issue of hotspots. Beyond that, it could help us both to get at a better understanding of the possible mechanisms behind rearrangements, and to identify which have indeed happened. This in turn could help improve the models for comparing genomes, deriving possible ancestors and, ultimately, understanding the course of evolution and its functional impact.

The fine analysis of the regions around breakpoints with the aim of better understanding the underlying mechanisms that have lead to the breaks will thus be the second main concern of this survey. To simplify matters, we shall concentrate our attention on the evolution of mammals only. The genomes of other species are as mobile but probably present a different dynamic.

Much is already at least partially known about the rearrangements that are possible and about their underlying mechanisms, both from a "pure biology" point of view, and through some initial computational studies that were done in the past, at a small or larger scale, and are appearing with an increasing frequency. Much more is not known. Finding one's way in the written or oral literature to piece all the information together, or just to precisely identify what is and what is not known, is, however, like trying to find a set of needles dispersed inside thousands of haystacks.

This paper will therefore have the pretence only to serve as an initial kick into the investigation.

[1]Cytogenetics is the branch of biology that deals with heredity and the cellular components, particularly chromosomes, associated with heredity.

We hope that it helps at least to show that the issue is even more complicated than already thought, and far more fascinating.

The paper is organised as follows. We start by giving a general introduction to genome dynamics, including quickly presenting some of the experimental techniques used to identify and study rearrangements. We then explore and discuss the methods developed with the purpose of detecting breakpoints (or their close cousins, conserved segments), and those methods that could be hijacked to do that. We then get to the heart of this paper, which is a survey of what is known, through genomic approaches, about possible rearrangement mechanisms. We end with a general discussion and some open questions.

2 Biological background

The expression "genome dynamics" refers to the structural variations observed in genomes along the course of evolution. Besides punctual mutations, genomes thus undergo large scale changes that have been called *rearrangements*. These involve parts of the genome that may be of varying length, from several kilobases to entire chromosomes. Several types of rearrangements are also to be distinguished: inversion, duplication and deletion of a segment inside a chromosome, transposition, reciprocal translocation which is the exchange of two segments between two chromosomes, fission which is the breakage of one chromosome in two and fusion of two chromosomes, that is, their joining into one. Such rearrangements play an important role in evolution and speciation although it has been observed [**?**] that not all rearrangements create a species barrier as was originally believed. Indeed genomic structural polymorphisms have been observed in individuals within a same population [**?**]. Rearrangements have however often been associated with genomic disorders [**?**] and have therefore been well studied by biologists for a long time, with the aim in particular of understanding their underlying molecular processes.

It is largely accepted that most rearrangements are initiated by one or several Double Strand Break(s) (henceforth denoted by *DSB*). A DSB is a break that cuts at a same position the two strands of a DNA molecule, as opposed to a Single Strand Break. Such lesions are not rare and may be induced by various factors, for instance, by Reactive Oxygen Species (oxygen ions, free radicals and peroxides), ionizing radiation (X and gamma rays), replication across a nick, and so on. In some specific cases, DSBs may also happen in a voluntary and programmed manner, taking part in a more complex molecular mechanism process, for instance DSBs may be generated by specific enzymes during V(D)J recombination in lymphocytes, or in the crossing-over process during meiosis.

A V(D)J recombination is a process that participates in immune cell protection. It generates variablity in the immune response molecules, which is essential for the cell because it enables the recognition of a great number of "foreign" entities in the organism. Starting from an original set of DNA segments, a V(D)J recombination generates different combinations among them thanks to site-specific DSBs. Recombination in general, not just V(D)J, is a complex and well documented mechanism in molecular biology. It allows the exchange of DNA segments between two DNA molecules (or two parts of a single molecule). The process is initiated by nucleotide pairing between the two molecules, thus a stretch of sequence similarity is needed. Then the two molecules are intertwined, and this leads to a complex conformation called the Holliday junction, which can be resolved by the exchange of segments between the two molecules. As a simplified example, a recombination between two molecules AB and A'C at a locus A can lead to the molecules AC and A'B. As concerns meiosis (gamete generation step), DSBs may be involved in it through genetic recombination (or crossing over). The latter plays a crucial role in the generation and maintenance

of genetic diversity by shuffling alleles between homologous chromosomes.

In the above cases, DSBs appear useful for the cell but they are always a source of serious damage if they are not repaired because the genomic integrity of the cell is endangered. Indeed, a single DSB may be sufficient to stop the cell cycle. Contrary to Single Strand Breaks that can easily be repaired using as template the unbroken strand, the repair of a DSB requires more complex molecular mechanisms. At least two such repair processes are largely described in the literature. They are called Non Homologous End Joining (denoted by *NHEJ*) and Homologous Recombination (*HR*). The first one is a *biochemical* process in the sense that the repair is done regardless of the initial DNA information. It consists in joining the two broken ends, but this costs the loss of some parts of the DNA molecule at both extremities [**?**].. On the contrary, the second mechanism, Homologous Recombination, is more conservative. We may call it a *genetic* mechanism: it restores the injured genetic information by using a similar one. This similar genetic information comes from the chromosome homologous to the one broken that is thus employed as a template to repair the broken chromosome. This repair mechanism is based on a recombination process [**?**]. The main difference between NHEJ and HR is that HR requires long stretches of similarity while NHEJ may involve sequence similarity but of shorter sequences. This is the reason why NHEJ is also called non-homologous recombination. The choice between the two repair mechanisms seems clearly determined: it depends on the DSB origin and on the state of the cell relatively to the cell cycle [**?**, **?**].

Rearrangements occur when the repair mechanism fails, or when it makes a mistake. NHEJ is likely to misrepair when two (or more) DSBs occur simultaneously on a genome. The joining of broken ends that do not come from the same breakpoint would in this case generate a rearrangement. For instance, if two DSBs occur on the same chromosome and the NHEJ misjoins the broken ends, then the segment between the DSBs will be reversed. The HR mechanism can err if a wrong template is used. Indeed, only sequence similarity is needed to initiate the recombination, and if the sequence used as template is not orthologous (coming from a same ancestor through speciation), more precisely if it is not at the same locus on the homologous chromosome, it may generate a rearrangement. Indeed if the recombination occurs between orthologous loci (allelic recombination), an exchange of DNA may happen but the genomic organisation will not be altered since the localisation of the exchanged material does not change on the chromosome, whereas recombination between different loci will lead to changes in genomic organisation: the exchanged material will no longer be at the original loci. This latter process is called Non-Allelic Homologous Recombination (*NAHR*). It has been shown to be the mechanism responsible for several human genomic disorders (reviewed in [**?**, **?**, **?**]), particularly when it leads to what is called unbalanced rearrangement, which is a rearrangement leading to the gain or loss of DNA. Contrary to HR, NHEJ, as far as we know, has rarely been implicated in evolutionary or disease rearrangements. We assume the reason is that this mechanism has left no trace of its occurrence (or none yet detected). Nevertheless, it is generally admitted that NHEJ can generate rearrangements; for instance [**?**] has experimentally determined the frequency of translocations generated by NHEJ (less than 3%).

Finally, to be viable, a rearrangement must sustain the different steps of the cell cycle, such as replication and mitosis. Moreover, to be transmitted to the offsprings, a rearrangement has to occur in the germ line, and to successfully pass the meiosis step. This is a biologically delicate and difficult step because meiosis can complete only if the chromosomes are correctly segregated. It is known, for instance, that some translocations are not possible because they prevent the right segregation of chromosomes [**?**]. Further, a rearrangement has to be selected and fixed in the population, which means that it has to provide some selective advantage. A rearrangement may also be polymorphic. Polymorphism is a condition in which a population possesses more than one allele at a locus. There may be several causes why polymorphic rearrangements is observed. For

instance, they can be maintained by a balance between variation and natural selection, or because some heterozygous advantage is conferred over individuals who have two copies of the wild type allele. If selection is operating, migration can also introduce polymorphism into a population. Multiple niche polymorphism exists when different genotypes should have different fitnesses in different niches. Genetic drift is a further possible source of genetic variation.

3 Detecting breakpoints

To study rearrangements, the only data available are the actual arrangements of the genomes. To reconstruct the rearrangement scenarios that have occurred since the divergence of two genomes, the first step is to identify the regions of the genomes that have not been broken, that is, the conserved segments. We may assume, by the parsimony hypothesis, that they must derive from the same region in the genome of their closest common ancestor.

3.1 Experimental methods

Various experimental methods have been developed to analyse karyotypes and identify conserved segments. A karyotype is the complete set of all chromosomes of a cell of any living organism. It is a screenshot of the chromosomes of a genome.

The first, and most intuitive, approach developed was to compare karyotypes of several organisms or individuals. By this means, it was possible only to "see" the differences in number or in size of chromosomes. Then, in the 1970s, a technique called chromosome banding appeared that enabled the identification of rearrangements at a finer scale. Using some coloration solutions, this technique allows the differentiation of several kinds of bands on chromosomes (the size of a band is roughly 4 Mb). Therefore, a chromosome can be characterised by its band pattern. This allows the comparison of the karyotypes from different species based on such patterns. The resolution remains however low, and only major rearrangements can be identified with this method, like chromosome fusions or fissions, large translocations and inversions. Moreover, the bands can be misleading when the considered species are not closely related because the assignment of homologous bands becomes otherwise too difficult.

Then in the 1990s, a major technique in the field of cytogenetics was developed based on the principle of hybridization: Fluorescent In Situ Hybridization (FISH) [**?**, **?**]. Briefly, hybridization is a molecular process that joins two complementary single strand DNA molecules to form one double strand DNA molecule. FISH uses this process to locate probes, that is single strand DNA segments marked by fluorescence, on targeted sequences. For instance, FISH allows the detection of all the chromosomes of one species that share at least one conserved segment with a specific chromosome of another species. The resolution remains low because the fluorescent signal can not be detected if the fluorescence does not cover a sufficiently long sequence. Depending on the condensation level of the chromatine analysed, resolution varies from 50 kb (chromosome at interphase) to 3 Mb (chromosome at metaphase). This method, called comparative chromosome painting, and further extended to deal with more distantly related species (zoo-FISH), allows the identification mainly of inter-chromosomal rearrangements. It has been used to identify such rearrangements between a great number of species [**?**, **?**].

Array comparative genomic hybridization (array CGH, also denoted by array-based CGH) is another more recently developed molecular-cytogenetic method that allows to detect some types of chromosomal changes, unbalanced ones only, not balanced reciprocal translocations nor inversions. In particular, it is being extensively used to analyse copy number changes (losses, gains and amplifications) in the DNA content of cells. The technique is derived from conventional CGH, which

enables to characterise both somatic and constitutional genomic DNA mutations. In conventional CGH, the DNA of interest and a reference are fluorescently labelled and hybridized to a normal metaphase preparation. In array-based CGH, large insert clones like BACs and PACs, containing human DNA, replace the metaphase preparation as target. Using microscopy and quantitative image analysis, regional differences in the fluorescence ratio of the DNA of interest versus the reference can be detected and used for identifying abnormal regions in the first. Both CGH and array CGH do not however provide information as to the precise location of the rearranged sequences.

Another experimental technique, called gene mapping, is also used to study rearrangements. It preceded FISH or genome sequencing. The aim is to experimentally locate the genes on a genome with respect to one another. Gene mapping is usually done by two main techniques: linkage analysis and radiation hybrid mapping. The idea of the former is that if two loci are "linked", they are inherited together. The relative distance between two genes may thus be approximated by estimating the frequency at which they are observed to be simultaneously inherited, assuming that the distribution of crossing overs is uniform along the genomes. Radiation hybrid maps are obtained by irradiation of the studied genome before fusion in other cell lines. The irradiation cuts the chromosome in different fragments, which are independently kept or lost during the cell life (culture). The distance between two markers can then be estimated using the frequency at which they are found together, assuming that the closer the markers are, the less they are separated by irradiation.

Studying rearrangements using gene mapping data, consists then in comparing gene orders. Of course, only those rearrangements that involve a gene can thus be studied, and the technique requires a good identification of the genes, and of the orthologs between species.

Finally, another novel technique appeared recently, called "end-sequencing profiles" (ESP for short). It consists in cloning BAC-sized parts of the genome of interest and sequencing only the extremities. The latter are mapped on a reference genome, mainly using sequence alignment. The space between two corresponding ends are then compared with the normal size of a BAC sequence. If the distance is too different, it means that at least one rearrangement distinguishes the genome of interest from the reference one. This technique is mainly used on cancer or polymorphism data because it is cheaper than sequencing the whole genome of onterest. It does however require that the reference genome is wholly sequenced.

In this survey, from now on, we concentrate only on the genomic methods to identify conserved segments. Such methods are based on the alignment of whole genomes. This enables the identification of the conserved segments between two genomes at a finer scale than using FISH or similar methods which can in general not detect bands smaller than a few megabases in size. FISH does not allow either to detect intrachromosomal rearrangements. Genomic methods are also more precise than gene mapping which further relies on orthologous assignations that are often error-prone. However, it presents the inconvenience of being applicable only when genomes have been wholly sequenced, and there are less of those than of genomes to which FISH-like techniques have been applied. Furthermore, comparing genomic sequences is not a trivial problem. For instance, whereas hybrids obtained by FISH may be considered (directly) as true homologs, like in the case of gene mapping, genomic alignments do not allow for a fully reliable identification of homology, or worse, of orthology. The ideal then, as argued in [**?**, **?**, **?**], would be to use both types of data simultaneously, something that has rarely been done up to now.

3.2 Genomic alignment

There are two main types of alignment algorithms: local and global. Global alignment algorithms, first described in [**?**, **?**], seek to align two sequences from the beginning to the end of each. This

therefore requires the two sequences to be well conserved with no changes in the order and orientation of any of their segments. On the other hand, local alignment algorithms, the first of which is due to Smith and Waterman [**?**] find the segments of each sequence which obtain the best alignment scores when aligned to each other. It usually outputs a set of such maximal scoring, in general non overlapping segments. Working with whole genomes, and in particular with vertebrate genomes, adds further difficulties which prevent us from directly using either of those classical algorithms.

First, the size in base pairs of the genomes makes it impossible in practice to align them using an exact algorithm. For instance, the human sequence assembly contains roughly three billion base pairs. Heuristics have therefore in general to be used, and preferably very fast ones.

Second, genomes are a more extreme case of sequences that can not be aligned globally. As an example, human and mouse have diverged seventy five million years ago and, according to some sources, only forty percent of their genomes can be aligned [**?**]. Indeed only a proportion of the genome is under selective pressure (5% is the estimated value from a comparison between the human and mouse genomes [**?**]) and thus well enough conserved. The coding regions are believed to represent only a small percentage of the genome (roughly 1.5% [**?**]) while most alignment algorithms were designed precisely for coding sequences. Aligning intergenic sequences is much more difficult because they are less conserved and can include segments with no detectable similarity to their orthologs in other species. Such segments therefore can not be aligned well with any currently known technique.

Finally, genomes may have undergone rearrangements, which is the initial motivation for aligning them, at least in this paper! Duplicated elements for instance may be a problem because, even though some of them can be masked before the alignment step (for example, known transposable elements), a large number remains undetected, such as processed pseudogenes, duplicated genes, or segmental duplications. It is estimated that, for instance, roughly 50% of the human genome is composed of transposable elements [**?**] and 10% of satellites (which are contiguous stretches of short to medium length repeats also called tandem repeats). Duplicated segments considerably increase the difficulty of identifying the segments that are orthologous (*i.e.* have kept the same ancestral position). In general, global alignment is not suited for whole genome alignment where rearrangements have occurred because the order and orientation of the segments is not conserved.

In the latter case, we can however hope that inside the conserved segments, sequences can be aligned globally. One possible strategy could thus be to divide the task in two steps: first find the most conserved parts (which will be called from now on *anchors*), and then align around the anchors to try to extend the (rearrangement-)conserved segments. This has been called the "seed and extend" method. To find the anchors, a fast all-versus-all local genomic alignment must be performed to get all the pairs of substrings which are, hopefully, homologous. The major difficulty of this step is to discriminate among the output hits between those which are true homologs (or, better, orthologs), and those which have occurred by chance. An intermediate step is thus needed to filter the anchors.

The strategy that is in general adopted by all current methods consists therefore in: 1. detect potential regions of homology (the anchors), 2. filter them (that is, eliminate false positives and make a choice between the different copies of duplicated elements), and 3. align (allowing for long gaps) the detected homologous regions. The first step (anchoring) is common to all methods while the last two steps differ among those that we describe next depending on what is the final goal of the method, "just" alignment or alignment with the purpose of studying genome dynamics.

Numerous methods have been developed, but we shall discuss mainly four of them. They were published roughly at the same time, *a.k.a.* when the mouse genome sequence became available (December 2002). The four main methods we discuss are: the GRIMM-SYNTENY algorithm [**?**], CHAINNET [**?**], an algorithm due to Couronne and Pachter among others [**?**] that we shall

henceforth denote by CP, and finally MAUVE [**?**]. The last one is quite different from the other three because it was designed mainly for bacterial genomes which contain a much higher proportion of coding regions (for example 90% in *E. coli*) and are therefore easier to align globally. All methods have been extended to deal with more than two genomic sequences (and therefore species), but we only discuss the case of two genomes. Only the GRIMM-SYNTENY algorithm was elaborated with the aim of studying genome dynamics. The motivations for the others are varied. An important but not exclusive one is to detect cis-regulatory sequences. The other methods we do not mention in this paper are either very similar in one way or another to the ones we detail, or are less appropriate for the purpose we have in mind (genomic alignments to detect breakpoint regions), or yet are not clearly enough presented that we may feel confident we fully understood the underlying algorithm (this is the case of the method used for the mouse genome [**?**]). However, we shall discuss briefly one of those other methods, SHUFFLE-LAGAN, later in the paper.

3.2.1 Anchoring

Local similarity between two genomes is first detected during the anchoring step. This is a step that is common to all methods. However, each uses a different model and algorithm for the task. Anchors may thus correspond to exact matches, almost-exact matches, or ungapped/gapped local alignments (that is, longer non exact matches that may also contain insertions and deletions). Anchors may be of varying length, and other restrictions may also be applied such as uniqueness, non-overlappingness etc. Concerning the four algorithms we discuss in more detail, all use ungapped (CHAINNET) or gapped (GRIMM-SYNTENY and CP) local alignments as anchors except MAUVE which works with exact matches.

In fact, CHAINNET uses a local alignment algorithm, called BLASTZ [**?**], for establishing its anchors. The algorithm produces gapped alignments, but CHAINNET then keeps only the parts containing no gaps. GRIMM-SYNTENY uses as anchors the gapped alignments given by the PATTERN HUNTER program [**?**], restricting those selected to a set of non-overlapping and unique ones.

The two methods have anchor finding algorithms that are quite similar. Both are based on a "seed-and-extend" strategy similar to the one used by the popular algorithm BLAST, that is, seeds (that correspond to short matches) are first extended without gap; if they score above a certain threshold, they are retained and further extended, this time allowing for gaps, and only those extensions scoring above a second threshold are kept. BLASTZ and PATTERN HUNTER differ mainly in the type of seeds they seek in the first step. PATTERN HUNTER introduced a novel type called "spaced-seeds" which consists in words of size l requiring matching bases only on a subset of the positions. BLASTZ has added to this type of seed the possibility of finding one transition (*i.e.* a mismatch A-G or C-T) instead of a match in one of the positions of the seed. Spaced-seeds have been shown to be very sensitive [**?**] and a benchmark study of seeds revealed that BLASTZ is one of the best seeded alignment algorithms for non-coding DNA because it allows for transitions [**?**]. This may be the reason why a more recent version of GRIMM-SYNTENY has changed the algorithm adopted for identifying anchors: this is now done with BLASTZ as for CHAINNET [**?**].

To find anchors, CP uses the local aligner BLAT [**?**] which is also based on a "seed-and-extend" strategy. The seeds are in this case exact or almost-exact matches, and the originality of the algorithm is to group the seeds which are close together on the same diagonal (that is, stand at a same distance from one another in both sequences), and to extend only those present in a significatively big group. Because it is less flexible in the definition of anchors, BLAT was not designed for cross-species alignment but it presents the advantage of being very fast.

The above anchoring algorithms require several parameters to fix, such as the length and other characteristics of the seeds, extension thresholds, substitution matrix and gap penalties. The choice of the parameters depends on the conservation rate of the aligned sequences (therefore on the species considered), and on the desired trade-off between sensitivity and specificity. Often this trade-off is determined "with the hands", and is not well described.

MAUVE is different from the other methods in the anchoring step because it is very stringent: it retains only exact matches, and in addition such matches must be unique. This anchoring step is less sensitive than gapped alignment algorithms and MAUVE seems therefore not well suited for distantly related species.

The sizes of the anchors found by those different methods vary from a few bases (MAUVE), to longer segments (around 500 base pairs for GRIMM-SYNTENY with a more intermediate value of 30 base pairs for CHAINNET).

More recently [**?**], it was suggested that for aligning evolutionarily more distant species, anchors restricted to the coding parts of a genome would be more appropriate. Indeed, Bourque *et al.* tried to compare mammalian genomes with chicken. Since the two sets of species are more distant, intergenic DNA is less conserved. One must therefore use very sensitive parameters for the alignment, but this then increases the rate of false positive hits. On the other hand, coding DNA is known to be more conserved among species, and the alignment can be done at the protein level, which is more specific than at the nucleotide level. The anchors in this case are then genes. The rate of false positives can be greatly reduced by adopting this strategy, even though mistakes may still be done, in particular with large gene families. Indeed, it is often difficult to discriminate between orthologs and paralogs. A filtering step is thus needed for this type of anchors. Moreover the orthologous gene sets obtained are smaller than the sets of "classical" DNA anchors produced by previous methods using the whole genomic sequences. The "gene-as-anchors" method prevents also the detection of purely intergenic conserved segments. However, Bourque *et al.* compared the results obtained with the two types of anchors and found that they are consistent [**?**]. On the website of Pachter[2], the CP strategy has also been updated to a gene-based anchoring method. In fact, CP now uses known and predicted exons (instead of genes) to anchor the global alignments. This tends to indicate that gene(exon)-based approaches are relevant.

3.2.2 Clustering or chaining of anchors

The main idea in order to filter out false positive hits and retain only the true homologs is to use contextual information. For instance, two anchors occurring near to each other and at a same distance in both genomes are less likely to be due to chance than a single isolated anchor. How precisely such contextual information is technically taken into account represents a crucial step for an accurate local genomic alignment. Two types of approaches have been considered to filter anchors that preserve the flavour of the above example while allowing for more flexibility: one consists in chaining the anchors, the other in clustering them. The only criterion used by the latter is the relative distances between anchors, whereas chaining requires also the anchors to appear in the same order and orientation in the two genomic sequences. Both GRIMM-SYNTENY and CP filter anchors by clustering them while the two other methods we discuss (MAUVE and CHAINNET) use chaining to try and eliminate false positives.

In GRIMM-SYNTENY, the distance between two anchors is defined by the Manhattan distance: it is the sum of the distances in the two genomes. An anchor is added to a cluster if its Manhattan distance with (at least) one of the anchors already in the cluster is less than a specified threshold.

[2] http://math.berkeley.edu/~lpachter/

In order to be retained, a cluster must span at least C base pairs where C is a parameter of the algorithm and represents some positive integer. One possible drawback of this method is that one may obtain overlapping clusters even though the anchors themselves can not overlap.

The strategy of CP, as far we understood it (the paper is not perfectly clear on this point), neglects, like GRIMM-SYNTENY, the order and orientation of the anchors during the filtering step, but the last step consists in a global alignment of the clusters. Clusters containing too much disorder will then probably not lead to a significant global alignment score and will eventually be removed. It is not clear either what is the criterion adopted for clustering, although this seems based, as for GRIMM-SYNTENY, on a distance, although perhaps not a Manhattan distance, between successive anchors.

On the other hand, the two remaining algorithms we consider for discussion take (directly) into account the order and orientation of the anchors. MAUVE does not use the information of distance between successive pairs of anchors, but retains only the chains that, once aligned, cover enough base pairs. The chains can not be intertwined with one another, since the order constraint is strict and precludes the existence of intervening blocks in a chain. CHAINNET makes chains of anchors using a k-dimensional tree structure [**?**]. This is a space-partitioning data structure for organising points in a k-dimensional space. It can be used, for instance, to find all the points that lie within a given rectangle or higher dimensional space. Contrary to MAUVE, CHAINNET allows chains to overlap and does not discard any chain, but assigns a score to each (depending on the number of anchors it contains, their spacing etc.). We shall see in the next section how CHAINNET then uses this list of scored chains for the last step of the method.

When dealing with genes or exons as anchors, the latter are clustered or chained based on their order only, except in the case of Bourque *et al.* [**?**], who continue using a Manhattan distance. The unit for calculating such a distance between anchors is not a base pair anymore but a gene.

3.2.3 Extension or recursivity

Once sets of anchors have been selected (and some false positives hopefully eliminated), whether by applying a clustering or a chaining method, they must be used to produce a final alignment of the genomic sequences. This may not lead to a global alignment but only to extended local alignments around the clusters or chains identified in the previous step.

As mentioned, this last step is the most different among the methods that we have been considering, mainly because theirs objectives are also different. GRIMM-SYNTENY is the only one that was designed specifically and exclusively for the study of rearrangements (CHAINNET was also used for the same purpose but not only). GRIMM-SYNTENY therefore does not attempt to perform a global alignment, nor even to precisely align the conserved segments. Moreover, such conserved segments, called by the authors *synteny blocks* (we shall use both expressions indifferently from now on), are used to reconstruct the history of the rearrangements using permutation algorithms. GRIMM-SYNTENY therefore aims at obtaining long synteny blocks that include perhaps some small rearranged parts called *micro-rearrangements*. With this purpose in mind, GRIMM-SYNTENY tries to form strips of clusters. The latter are clusters that appear in the same order and orientation in both genomes although they can stand far from one another. It remains however unclear for the authors of this paper after reading the appendix of [**?**] where it is best explained that the algorithm for determining the orientation of a cluster is correct in all cases. For instance, to the best of our understanding, it seems enough that the first anchor in a cluster is not rearranged in relation to the other anchors for the cluster to be assigned a positive orientation, no matter how rearranged may be the remaining anchors. The orientation of a cluster is clearly not a trivial problem that, to our understanding, is not yet fully solved.

On the other hand, the objective of CP was not to study rearrangements but to align two genomes. Therefore, it does not try to extend the clusters obtained at the second step, but just to get them aligned. It uses for this a global aligner, called AVID [**?**], that is applied to each cluster. Only the alignments scoring above a fixed threshold are retained and output.

The idea behind CHAINNET is rather original. As long as some sequence positions are not covered, the algorithm greedily selects the chains previously obtained by decreasing order of score, marking at each step the sequence positions that are covered by the latest chain selected. Since the chains of CHAINNET may overlap, contrary to what happens with MAUVE, a same position may be covered by more than one chain. CHAINNET seems to discard positions of a chain that are already covered by a previously selected chain, but it is not clear whether and how it then keeps track of such positions to help identify potentially duplicated regions. It probably registers only that a position is part of a possible duplication somewhere else in the geome. CHAINNET then outputs a net of chains, since a chain which is embedded within a gap of a previously selected chain is labelled as being a child of the latter. However, it is again not clear how to use such a net, and whether different levels of the net have a different biological meaning.

Finally, MAUVE repeats the two first steps (anchor finding and chaining) with less stringent parameters inside and outside the chains of anchors previously obtained, and not overlapping with those. Chains selected at some step are never questioned. When no new chains exist, the regions not yet covered by a chain are aligned using a classical alignment algorithm.

3.2.4 General comments

We would like now to spend a few minutes commenting on the four methods presented above. Our purpose is not to detail the advantages and inconvenients of each but only to discuss them in the light of what was our initial purpose for getting interested in them. We remind that this initial purpose is to consider methods that have been developed, or could be deviated from their initial objective in order to precisely detect the points along a genome where a break has potentially occurred.

With this aim in mind, there are clearly two issues that are of main interest for us. These are how the different methods deal with micro-rearrangements, and how they deal with duplications.

Concerning the first issue, the application of GRIMM-SYNTENY will have the effect of masking the presence of micro-rearrangements, and in such a way that it would be difficult to recover them *a posteriori*. The argument used by the authors for justifying this procedure is that a great number of such micro-rearrangements may be the result of an assembly error. This may be debated and in the case where it is crucial to find **all** breakpoints, this might not be a method of choice to adopt. For some purposes however, it may be less important to miss some breakpoints. Of course, whether breakpoints are indeed missed by GRIMM-SYNTENY depends on how parameter C of the algorithm is set. However, if augmenting the resolution may enable to deal more appropriately with the problem of masked micro-rearrangements, it may also increase the rate of false homologous assignations.

The small conserved segments, and therefore the breaks around them, may be relatively more easily recovered from the greedy process of chain selection of CHAINNET.

Finally, neither MAUVE nor CP address the question of micro-rearrangements but it seems that, like the other two methods, they will mask some of them during the clustering stage although this is harder to say as concerns MAUVE.

It is worth at this point mentioning an algorithm that was developed with the specific aim of aligning genomic sequences that have been subjected to small disruptions in the order of their segments. This is the SHUFFLE-LAGAN algorithm [**?**], that itself is an extension of LAGAN [**?**].

LAGAN is a global aligner that, like the other methods described here, first finds anchors which correspond to ungapped local alignments (using the CHAOS algorithm for that purpose), then retains a unique maximal scoring subset of such anchors, and finally performs a classical Needleman-Wunsch [**?**] algorithm to a narrow band around the anchors retained. SHUFFLE-LAGAN allows for some small local and independent inversions in the single chain of anchors it keeps for the final so-called *glocal* alignment. This would seem to make of it an algorithm of choice for the purpose of detecting the breakpoint regions. The problem is that a unique maximal chain is kept, and that in this unique chain small local inversions only are permitted. This precludes the direct application of the method to detect breakpoints in genomes which have suffered other types of rearrangements and/or large ones. However, it could replace a global aligner in the third step of the CP strategy for example; it would align the putative homology regions detected at the second step allowing for some micro-rearrangements.

Concerning duplications now, both GRIMM-SYNTENY and MAUVE will eliminate them, both because they deal with unique anchors. GRIMM-SYNTENY further disallows overlapping anchors. Such criteria will eliminate numerous useful anchors, whereas the filtering step might have enabled to pick the good orthologous copies. It is not completely clear what CP does with duplicated segments while CHAINNET apparently allows for such segments since successively selected chains may overlap in some of their positions (but it does not seem to keep track of which segments are the possible duplicates of another). SHUFFLE-LAGAN also deals with duplications but only if they are in tandem and present in only one of the two sequences. In both cases, this may not be the most satisfying way of considering duplications and may pose a problem in whole genome alignments, in particular for organisms that contain a high number of such duplicated regions, often very long, such as is the case for vertebrates, plants etc. However, the main difficulty comes from the fact that both CHAINNET and SHUFFLE-LAGAN introduce an assymetry in the way they deal with duplications: the latter are considered in one of the two genomes only. The main reason for this, or for choosing to eliminate duplications in some methods, including GRIMM-SYNTENY that was elaborated specifically for studying genome rearrangements, seems to be algorithmic complexity, although some types of duplications have started being considered by methods for computing permutation distances between genomes based on gene order data [**?**, **?**, **?**].

In general, all the methods for aligning sequences present problems: subjective thresholds or parameter values at different steps of the algorithm, possible dependence on the order with which some steps are applied, too little or too much sensitivity. Some of them have been compared, often briefly, among them. It has thus been "shown" [**?**] that GRIMM-SYNTENY and CHAINNET obtain roughly the same final "conserved segments". What usually differs among the various methods is the size and number of the resulting synteny blocks as these depend on the user-defined value for the parameters. From a preliminary examination of those methods, we have the impression, which should be further verified, that the crucial step is the filtering one, and that most methods are robust to changes in the anchoring step, assuming it is sufficiently sensitive.

Table 1 presents a summary of the main characteristics of all the methods commented upon in this paper.

Finally, an application of GRIMM-SYNTENY to human and mouse revealed 245 macro-rearrangements and 3170 micro-rearrangements inside the synteny blocks. CHAINNET applied to the same genomes detected 160 inversions of more than 100 kb, 29 closely located duplications or translocations and 46 distant duplications or translocations. It is important to observe that such numbers can not be directly compared because different parameters are used by the algorithms, and, more crucially, because different types of rearrangements are considered by each. CHAINNET only detects embedded inversions and non-overlapping rearrangements, whereas GRIMM-SYNTENY reconstructs a rearrangement scenario from the conserved segments and can deal with mosaics (overlapping) of rearrangements. Although the latter is more realistic, it is important to observe that it remains imperfect because it implicitely or explicitely assumes a certain relative

Method (Author)	Length of considered blocks	Genome coverage	% of aligned base pair	Number of synteny blocks	Average size of blocks	Inter-block average size
CHAINNET [**?**]	> 100 kb	90.9 %	32.9%	579	983 kb	450 kb*
GRIMM-SYNTENY [**?**]	> 1 Mb	93%	?	281	9.6 Mb	668 kb
CP [**?**] [1]	> 100 kb	76%	< 35.2 %	8080	270 kb*	86 kb *
exon-based map (unpublished)[1]	?	80%	?	494	4.76 Mb	?

Table 1: Comparison of the blocks obtained with the different algorithms studied, between human and mouse (except for CP and the exon-based map which has been obtained from the comparison of human, mouse and rat). A question mark in one of the cells indicates that the information could not be found in the paper describing the method. A star indicates that the value has been computed (by us) using the following formulae. The inter-block average size is given by: $\simeq$ genome size $\times$ (1 $-$ coverage))/nb blocs. The block average size is given by: $\simeq$ genome size $\times$ coverage/nb blocs. [1] Data obtained from [**?**].

frequency of occurrence for each type of rearrangement. For instance, transposition events are modelled as three inversions.

One may wonder which of the methods we presented above is the most appropriate for the purpose of precisely analysing the breakpoint regions in mammalian genomes. We can not conclude on the MAUVE algorithm without further studies. Indeed, MAUVE was designed for bacterial genomes and because of the stringency of its anchoring step, is better suited for closely related species. It has been used only once (at least as reported in literature) on mammalian genomes, furthermore to detect major rearrangements [**?**]. The CP algorithm was not made for identifying breakpoint regions, or its dual, conserved segments. In fact, it outputs around 8000 blocks (see Table 1) without testing if some could be merged according to order and orientation. Finally, GRIMM-SYNTENY and CHAINNET identify conserved segments (or synteny blocks), and thus breakpoint regions. The latter however are long (of, on average, 668 and 450 kb respectively) making it more difficult to further analyse them in detail and precisely locate the break positions.

4 Analysis of breakpoint regions

We are now ready to address the question that lies at the heart of this paper. This concerns the state of our current knowledge on possible rearrangement mechanisms, particularly as this may be obtained through genomic approaches. The survey below is certainly far from representing the whole of this knowledge, even the knowledge derived solely from sequence analysis, but we hope it provides a fair enough view of it.

Two main kinds of breakpoint analysis have traditionally been done: the first investigates a specific breakpoint at a time (we henceforth call such type of analysis *punctual*) while the second corresponds to analyses which deal simultaneously with all, or almost all the detectable breakpoints in a genome (we call such type *systematic*). Historically, cytogeneticists have been concerned with the first kind of analysis, the main objective being to study a specific disease. The interest is thus in discovering the putative causes of the rearrangement(s) responsible for the disease and in understanding the underlying molecular mechanisms. Cytogeneticists have of course been also interested in *evolutionary* breakpoints, that is breakpoints which enable them to differentiate among species. The reason why their study of the latter used to be punctual was due simply to the limited amount of data available at the time. In fact, cytogenetic experiments generate large breakpoint regions, which must be refined further to more precisely delimit a smaller area around the break. The refinement is often done by conducting FISH-like experiments. Before the availability of the whole genome sequence of the studied organisms, cytogeneticists were obliged to sequence themselves the regions of interest, a process that was then slow and costly. The advantages of a punctual analysis of breakpoints are that they can be very detailed and may thus have more explanatory power than systematic studies. Their main inconvenience is that their very scale often precludes evaluating the statistical significance of the

observations made.

The whole genome sequencing projects of the last decade or so provided the data, and thus the opportunity to perform more exploratory and systematic analyses of the breakpoint regions. For the first time, it is thus possible to attempt finding common features among all detectable breakpoint regions. The objective, which remains the same, that is to understand the molecular mechanisms underlying rearrangements, was given a "boosting" motivation by the recent polemics on a "right" model for the appearance of breakpoints: do they occur at random positions as Nadeau and Taylor suggested in the 1980s [**?**], or are there hotspots along a genome where the breaks preferentially occur as argued by Pevzner, Kent and others [**?**, **?**]? However sequencing a whole genome is long and expensive process and thus such type of data is limited to evolutionary studies. That is the reason why polymorphism or disease analyses are mainly made using cytogenetic data.

We survey both types of analyses, starting with the systematic ones and ending with the punctual studies.

Before that, we would like to call attention to three general observations. The first is that when studying breakpoints, one should always keep in mind that their characteristics, distributions and underlying mechanisms may be different depending on whether the breakpoint is somatic (the associated rearrangement affects, at mitosis, only one cell, like in cancer, and is not transmissible to the descendants) or occurs in the germ line cells (this type of rearrangement occurs during meiosis and is inherited; this includes rearrangements involved in genomic disorders, in evolution and in polymorphism). It is generally accepted that somatic tumor breakpoints are not randomly localised [**?**]. The polemics we mentioned above on the random versus non-random distribution of breakpoints refer only to evolutionary ones. We shall also be concerned mainly with the latter type of breakpoints only.

The second observation when studying breakpoints is that the choice of species to examine is important. Closely related species provide in general more details for the analyses as the traces left by the rearrangements have not yet had time to be scrambled by further evolution. However, there are often less breakpoints between close species. One should be aware also that the mechanisms for rearrangements could be, at least in part, species-dependent [**?**]. In this survey, we focus mostly on studies that were done on mammals.

The last observation is not a trivial one, it addresses the question of the definition of a breakpoint, or breakpoint region. So far, we have considered a breakpoint region to be the genomic region in between the detected conserved segments. Its size depends therefore on the resolution of each method. From a biological point of view, a breakpoint is the portion of a genome that has been involved in a rearrangement process. However, since the molecular mechanisms underlying a rearrangement are not well known, it is often difficult to precisely define what this portion is. Moreover, its size may also depend on the type of rearrangement that took place. Based on our current state of knowledge of these mechanisms, it seems clear that more than one nucleotide is concerned, but how many more remains unknown. When looking for sequence characteristics linked with rearrangements, the question of the size of the regions to analyse is crucial, yet it is not discussed in the majority of papers.

4.1 Systematic studies

Systematic studies are probably meaningful only when they can be done on an unbiased subset of all the breakpoints. Although this may be as impossible to obtain as the whole set itself, there are certainly precautions that can, and must be taken. One may nowadays, for instance, more easily avoid the biases that were introduced by the early cytogenetics methods due to the limitations in resolution of the banding techniques. Indeed, the latter enabled the detection of major rearrangements only.

Most systematic studies use the alignment of whole genomes, as described in Section 3.2, to identify the breakpoint regions. The results obtained depend of course greatly on the initial data, on the definition of conserved segment, or synteny block adopted (whether it allows or not for micro-rearrangements and duplications), and on the strategy and resolution chosen for computing such segments (nucleotide or gene/exon-based).

4.1.1 Random or not random?

The first characteristic that was discovered when whole genomes started being aligned is the loss of similarity in-between the larger blocks of synteny. Authors differ however in their interpretation of this phe-

nomenon. Some assign it to alignment errors or artefacts [**?**], while others consider this as the result of micro-rearrangements that have shuffled small segments in the breakpoint regions [**?**]. The latter authors argue that such micro-rearrangements plead in favour of the non-random model for the distribution of breakpoints along a genome. Since numerous rearrangements occur in the same regions, these regions are more prone to rearrangements. Kent and colleagues thus detected 19800 short chains less than 100 kb at the top level (*i.e.* between the longer chains) of the net and many more at the lower level (intermingled within the longer chains).

The other argument, by Pevzner and colleagues [**?**, **?**] in defense of the non-random model, is based on a different observation. After having detected the blocks they identified as syntenic between human and mouse (using GRIMM-SYNTENY), Pevzner and colleagues applied a reconstruction scenario using the blocks as markers. The scenario took into consideration four types of rearrangements: inversions, translocations between chromosomes, fusions and fissions of chromosomes. The authors then noticed that some inter-marker regions are re-used, suggesting that these regions correspond to hotspots. The "re-use" issue is one of the most debated on the topic of hotpots [**?**, **?**, **?**]. Both Kent *et al.* [**?**] and Pevzner *et al.* [**?**] say however that if one looks only at the distribution of the lengths of the large synteny blocks each method identifies, then this distribution is in agreement with the random model of Nadeau and Taylor. It is the observation of breakpoint re-use for [**?**] and of the many short rearranged segments for [**?**] that, for them, militate in favour of the non-random model.

On the side of the random model, Sankoff and colleagues [**?**] tried to understand the reason for the loss of similarity between the blocks of synteny. They thus sought to analyse further the small segments found between these blocks. Taking the human genome as reference, they distinguished different types of segments depending on which chromosome of the mouse they map to: the same as the adjacent blocks of synteny, the same as at least one block that is relatively close in the same chromosome but not adjacent, or to altogether different chromosomes. They found that these segments were not randomly distributed in the breakpoint regions and that most of them could in fact have been attached to the adjacent blocks of synteny. If they were not, this was because of an artefact of the alignment method used (that requires strong similarity at both ends of the blocks and hence leads to misalignments at these ends). The remaining segments could have been produced by micro-rearrangements or other molecular processes, such as retrotransposition. They argue that in general it cannot be ascertained that the rate of micro-rearrangements is greater outside the blocks than it is inside. One can thus not conclude against the random breakage model. They suggest an alternative hypothesis to explain the loss of similarity in the breakpoint regions: regions where a breakpoint first occurred (by chance) became then more prone to rearrangement **afterwards** and started evolving more rapidly (due to a quadrivalent conformation adopted during meiosis when chromosomes are in an heterozygous state). The idea that regions that have been subjected once to a rearrangement event evolve faster is also advanced, in another context than the random versus non-random polemics, by [**?**]. We shall come back to this soon.

4.1.2 Segmental duplications

Leaving aside this polemic and coming back now to what is our main interest in this paper, namely the analysis of the regions around where breaks have occurred following a rearrangement, one type of element has indeed been explored in more detail to determine its possible correlation with such regions. This is duplicated elements, more precisely *segmental duplications*, also called *duplicons*. The motivation for this is their obvious role in non allelic recombination [**?**].

Segmental duplications are large duplications that share high similarity (more than 95% identity) among copies, and that have a small amount of copies in the genome. They are also called *Low Copy Repeats*. Their length must be larger than 1 to 15 kb depending on the definition adopted by different authors. Segmental duplications in the human genome have been studied recently once the complete genome sequence of human became available [**?**, **?**, **?**]. They represent 5.2% of the genome and this proportion appears to have grown recently in the primate lineage. Their distribution along the genome is not uniform: 35% are localised in subtelomeric and pericentromeric regions, they are often clustered but they are not found in tandem, and the mechanism(s) responsible for their dispersion throughout the genome is still unknown. They are suspected to be associated with rearrangements because of their high degree of similarity and their great length. They thus constitute good substrates for homologous recombination between copies. Homologous recombination

between two copies at different loci would lead to a rearrangement, what we have called (Section 2) Non Allelic Homologous Recombination (NAHR). In fact, segmental duplications had been found already previous to the whole sequencing of the human genome, in studies of numerous genomic disorders in which they are believed to be involved (for reviews, see [**?**, **?**, **?**, **?**]). The mechanism had then been clearly identified as being NAHR.

Different authors [**?**, **?**] have more precisely looked for segmental duplications in the evolutionary breakpoints detected in the human genome by comparison with the mouse genome. They have shown that segmental duplications are associated with evolutionary breakpoints. Armengol *et al.* [**?**] found that 53% of the breaks of synteny contain at least one duplicon in a window of 25 kb around the breakpoint, and they suggest a putative role for these duplicons in the rearrangement process. Bailey *et al.* [**?**] found also a significant association: 26.5% of the breakpoints contain one or more duplicons at least 10 kb in size. The results between the two studies differ due to the methods that were applied and the thresholds used for conserved segment length, duplication length and so on. Contrary to Armengol *et al.* however, Bailey *et al.* do not believe that segmental duplications are a direct cause of rearrangements. In fact, segmental duplications appeared recently, they are primate-specific and are found associated also with mouse-specific rearrangements. Thus duplications and rearrangements did not occur in the same lineage and can not be linked. Rather, Bailey *et al.* think that duplications and rearrangements occur in the same region independently because the region is "fragile". Indeed, we can consider duplications as one more type of rearrangement. Their co-localisation among them and with breaks that result from other types of rearrangements would thus be yet another argument in favour of the hotspots or "breakpoint re-use" model. Segmental duplications have been found associated with breakpoints of synteny in several other studies. Zody and colleages analysed human chromosome 17 and found that 74% of the segmental duplications appear in regions of evolutionary breakpoints [**?**]. In another study, they detected duplicons in 13 out of the 15 breakpoints on human chromosome 15 [**?**]. Murphy *et al.* analysed 40 breakpoints involved in primate-specific rearrangements, 98% of them contained a segmental duplication, and in 62% of the cases duplicons were found flanking the rearranged segment [**?**].

More recently, another analysis of segmental duplications has been performed in the rodent lineage. Armengol *et al.* [**?**] thus found duplicons in 60% of the mouse-rat breakpoints and this led them to conclude to an association between duplications and breakpoints. All these results reinforce our belief that this association indeed exists, not only in primates, but also in other mammalian genomes, but that the cause-effect link can not yet be confidently established (that is, we can not yet conclude whether segmental duplications are a cause or an effect of rearrangements, or if both are completely independent events).

4.1.3 Various duplicated elements

Other types of duplicated elements have been searched in breakpoint regions. In a systematic analysis of the breakpoints of synteny on human chromosome 19, Dehal *et al.* [**?**] observed significantly more L1 (LINE 1) and LTR (Long Terminal Repeat) elements in those regions. L1 and LTR are two types of transposable elements (also called interspersed repeats) of the retrotransposon type. Retrotransposons copy themselves and paste copies back into the genome in multiple places. Initially retrotransposons copy themselves to RNA (transcription) but, in addition to being translated, the RNA is copied into DNA by a reverse transcriptase (often coded by the transposon itself) and inserted back into the genome. The other main type of transposable elements are DNA transposons that move by cut and paste, rather than copy and paste, using the transposase enzyme.

LINE elements have been found to be involved in a great number of deletions observed between human and chimpanzee, and several models have been suggested to explain their role in such type of rearrangements [**?**]. In the nematode genome, interspersed repeats and gene family members have been found associated with breakpoints of transposition and translocation events [**?**]. Nevertheless, this characteristic seems anedoctal because not always detected.

4.1.4 Evolutionary rates

Various analyses indicate that rearrangements tend to be associated with higher rates of evolution. Navarro and Barton initiated a debate on this issue with the analysis they performed on the human and chimpanzee

genomes [**?**]. Human chromosomes which show large structural differences with chimpanzee (a total of 10 out of 23 pairs of so-called rearranged chromosomes), exhibit greater synonymous and non-synonymous substitution rates than co-linear chromosomes (the remaining 13 pairs of chromosomes that are not rearranged). These results have been contested [**?**, **?**] and contradictory results have been obtained with the same human-chimpanzee comparison, with Vallender *et al.* [**?**] arguing that differential evolutionary rates are due to other factors. Nevertheless, several other analyses were performed at a finer scale and have pointed to similar trends. Marques-Bonet and colleagues have thus shown that breakpoint regions between human and mouse exhibit significantly higher rates of synonymous and non-synonymous substitutions, and further, that these rates are the higher, the closer one gets to the breakpoint [**?**]. Armengol *et al.* published similar results on mouse-rat breakpoints [**?**]. To explain this tendency, Navarro and Barton suggest the following speciation model [**?**]. When a rearrangement occurs in an individual, it does not create immediately a genetic barrier, as was previously thought, but this event reduces recombination in the rearranged region. The gene flow is then also reduced and incompatible alleles can accumulate which will then later on generate a genetic barrier leading to speciation. This model therefore predicts higher rates of evolution near breakpoints. This speciation model remains polemical but without getting involved in the polemics, one may still retain this tendency as one characteristic of breakpoint regions in at least some cases.

4.1.5 Fragile sites

Another interesting and completely different characteristic has been found in evolutionary breakpoint regions. This refers to the so-called *fragile sites*. Fragile sites are sites on the chromosome that have the tendency to break in specific cell culture conditions. They **must not** be confused with regions that break following a rearrangement. Very little is known about the causes of this fragility. Fragile sites are usually divided into two classes: the rare and the common fragile sites. The first appear in a very low percentage of the population, whereas the latter are believed to be present in all individuals. Contrary to rare fragile sites which exhibit some sequence characteristics such as short repeats expansion, common fragile sites have not so far being associated with any sequence features. They tend to replicate late and exhibit some specific banding characteristics. Nothing so far allows to explain the fragility of such sites [**?**, **?**]. Fragile sites are known to be involved in cancer rearrangements and the genomic regions where they are located can thus be considered as potentially rearrangement-prone. Ruiz-Herrera and colleagues [**?**, **?**, **?**] studied the breakpoints detected in primates by FISH and G-banding experiments. They tried to correlate them with some indices of fragility of the chromosome: fragile sites, intrachromosomal telomeric sequences and sites affected by X irradiation. They found mainly that evolutionary breakpoints are co-localised with fragile sites: in [**?**] 80% of the breakpoints occur at a fragile site, in [**?**] the proportion is nine out of nineteen breakpoints. However, it is important to observe that those analyses were made with banding data. Does it therefore mean something if a fragile site is located at more than 1Mb of a breakpoint? Moreover, no model is proposed to explain this correlation nor the possible role of fragile sites (as cause or effect) in the rearrangement process.

4.1.6 Correlations with other types of breakpoints (polymorphism, disease such as cancer)

Several systematic analysis have been performed on other types of breakpoints: rearrangements involved in inherited disease, in cancer, or even rearrangements found to be polymorphic in the population. Genomic disorders and cancer rearrangements, as we mentioned, are not necessarly created by the same mechanisms (especially for cancer) as evolutionary rearrangements. However, they have been extensively studied and the results obtained could be tested on the evolutionary breakpoints to help understand evolution mechanisms. Moreover, these different types of breakpoints have often been found to be co-localised, suggesting either some common mechanisms, or common preferences for certain regions [**?**, **?**]. Besides segmental duplications and low copy repeats, which are well known to be involved in these different types of rearrangements, other unexpected characteristics have been found associated to the corresponding breakpoints. For example, Abeysinghe *et al.* [**?**] conducted a systematic analysis on 219 gross deletion and translocation breakpoints involved in inherited disease or cancer. The authors looked for over- (or under-) represented characteristics in these breakpoint regions, exploring base, di- and trinucleotide composition, repetitions and recombination-associated motifs. Recombination-associated motifs are motifs that can drive non-homologous recombination,

that is NHEJ, which means that recombination occurs without sequence similarity or with only very short similarity, as opposed to homologous recombination that requires long stretches of similarity (of the order of hundreds of base pairs). Recombination-associated motifs have often a biological function, for example site-specific for V(D)J recombination, or as sites of fixation for topoisomerases. However, they can drive recombination outside their normal function and thus potentially generate rearrangements. In Abeysinghe *et al.*'s study, it was observed that deletion breakpoints are A/T rich whereas translocation ones are G/C rich and exhibit some specific tri-nucleotides over-representation. A number of recombination-associated motifs seem over-represented in breakpoints, and alternative purine-pyrimidine and pyrimidine tracts are over-represented in deletions and translocations respectively. In [**?**], the authors looked for short repeated sequences (of several base pairs) near the breakpoints involved in disease and cancer. They found such type of sequences in more than 80% of the breakpoints. These short repeats can form secondary structures which can bring distant DNA segments close to each other, and thus facilitate non homologous recombination between them. This theory does however not explain the initiation of the process, that is, the break itself. Bacolla *et al.* [**?**] show that some sequences are more prone to a break because they can form non-B DNA conformations (B-DNA conformation is the most usual DNA conformation: it corresponds to a right-handed helix). For instance, alternative purine-pyrimidine tracts can lead to the Z-DNA form (left-handed helix). Bacolla *et al.* [**?**] performed a systematic analysis of 222 breakpoints involved in disease and cancer, and found that the frequency and distance to the breakpoint of purine or pyrimidine tracts are significantly greater than expected by chance. They also analysed 11 cases in more detail, and in all these cases they have been able to show a putative non-B DNA conformation.

4.2 Punctual studies

As we mentioned, systematic studies were not the premier way of analysing breakpoint regions. The first types of analyses that were done (for many years) dealt with one or two breakpoint regions at a time. Those studies are quite different from the systematic ones and have the advantage that, after a more detailed exploration of the breakpoint regions, they often propose a model or hypothesis to explain the rearrangement mechanism behind the observed break. Most had a medical interest in mind although some concerned also evolution only. We focus mainly on the analysis of evolutionary breakpoints.

The putative causes of chromosomal changes involved in genomic disorders (or which are polymorphic) are easier to detect because one disposes of the "ancestral" sequence, which in the case is the healthy sequence (considered as being the sequence right before the rearrangement). This is not the case for the evolutionary rearrangements in general, for which the chronology of the events is not always easy to reconstruct.

The latter has been established since the 1970's by chromosome banding and FISH experiments between numerous mammalian species. However, the molecular characterisation of such breakpoints was performed more recently only (beginning of the current century). It was thus shown that, for instance, the human and chimpanzee genomes differ by 9 large pericentric inversions and one fusion. Almost all of these breakpoints have then been analysed at the molecular level and the breakpoint region confined to at most a few hundred kilobases [**?**]. Most of the breakpoint analyses were performed on closely related species, particularly among primates, mainly because the sequences are then easier to compare and the breakpoints can be more precisely detected.

When one tries to compare and combine the numerous results that were obtained by such punctual analyses, various common trends appear. The first is that such combined results are often in agreement with the results of systematic studies, which is reassuring. As a matter of fact, the breakpoint characteristic most reported in the literature of punctual and systematic analyses is segmental duplications. In punctual studies, they have been found associated with six of the nine inversions between human and chimpanzee. Their involvement in the breaks is however not always the same. They have been unambigously proven to be the cause of the pericentric inversion on human chromosome 18, through Non-Allelic Homologous Recombination between two intra-chromosomal copies [**?**, **?**], and also in the inversion on chromosome 16 [**?**]. A segmental duplication of 250 kb has also probably mediated the translocation that enables to distinguish gorilla from human and chimpanzee [**?**]. In other cases however, even though segmental duplications were detected in the breakpoint regions, such duplications seemed not to be involved in the rearrangement process. For instance, human specific duplications have been detected in the breakpoints of an inversion which occurred in the chimpanzee lineage (human chromosome 15). The interpretation of Locke *et al.* [**?**] is that this region is a

rearrangement and duplication-prone one. In their analysis of the pericentric inversion that distinguishes human chromosome 12 from its chimpanzee homolog, Kehrer *et al.* [**?**] conclude that duplicons found in the breakpoint regions arrived simultaneously with the inversion, to repair the broken ends of the DNA. Finally, in most cases it remains an open question whether duplicons are involved in the rearrangement process [**?**].

Other features have been detected in individual breakpoint regions, they are summarized in Tables 2 and 3. We highlight some of them only. Small rearrangements or deletions have been noticed in several rearrangement analyses [**?**, **?**, **?**], which are again in agreement with the computationally conducted systematic studies. Interspersed repeats are often reported in the literature [**?**, **?**, **?**, **?**], and such elements were also found over-represented in the breakpoints of human chromosome 19 [**?**]. Some features, known for potentially generating breaks in DNA, are also detected in several other studies. These are: topoisomerase fixation sites [**?**], poly-purine (or -pyrimidine) and alternative purine-pyrimidine tracts [**?**, **?**] and short direct or inverted repeats [**?**, **?**, **?**]. Such characteristics are often anedoctal and without systematic analyses one can not conclude to a significant role for them in the rearrangement process.

4.3 Conclusion and open problems

As one can see, it is still hard in most cases to draw any solid conclusions from the numerous studies that have been done, except maybe at the anecdotal level in some cases. Many more aspects of the chromosome breaking process remain thus unsolved than have been satisfactorily answered.

One major characteristic of breakpoint regions, perhaps the only one, that seems to be relatively reliable because it has been recurrently found is segmental duplications. These have indeed been identified as being linked to rearrangements in both systematic and punctual studies. In various cases, punctual analyses have allowed to definitely incriminate them in the rearrangement process, as a substrate for non allelic homologous recombination. Nevertheless, in many other cases, segmental duplications have also been proven **not** to be responsible for the rearrangement. In such cases therefore, the cause and mechanism for the latter remains to be found by exploring a possible correlation of the breaks with interspersed repeats, recombinogenic motifs, fragile sites, and so on. What seems at least clear is that several mechanisms for the generation of rearrangements co-exist.

Concerning the random versus non random polemic for the distribution of breakpoints along a genome, it would be tempting, based on the various results obtained by different authors that we briefly surveyed above, to conclude in favour of a non random model, and thus to the existence of hotspots of rearrangements. However, the non randomness of the *distribution* of breaks does not necessarily imply the non randomness of the *apparition* of breaks. Indeed, we can not forget the selective process that follows the birth of rearrangements. If the same rearrangement is observed in different lineages, perhaps this is because this particular rearrangement brings a selective advantage, contrary to others which are deleterious. On the other hand, the simple fact of finding sequence characteristics (any sequence characteristics) in the regions of synteny breaks could be further indices in support of the hotspot hypothesis.

If one tried now to enumerate what would seem to be the most promising avenues to explore in the future to get closer to this desired characterisation of the breaks in a genome, the first would probably be to combine the power of punctual approaches, specially as these are based on cytogenetics data, with the more systematic analyses. There is also cytogenetics data on more species than genomes sequenced and in particular FISH data on about 20 species seem to indicate that some of the results obtained with genomic data may not be correct [**?**]. It might help also the investigation if one had a more accurate idea of how evolution and speciation proceed: in a hierarchical way as is believed at least for non bacterial and viruses genomes, or netlike even for vertebrates as is suggested by Dutrillaux [**?**]? As we saw, duplication events are in general not considered in the whole genome alignment methods that are used previous to a systematic analysis of breakpoint regions. It is obvious that they should be taken into account. Being able to distinguish the different types of rearrangements and conduct the analyses on the breakpoint regions for each type separately might also greatly help identify the characteristics, if any, of those regions. The existence of rearrangement polymorphisms, and the true relation between rearrangement and speciation would also be worth exploring as it may shed a better light on the evolutionary process. Indeed, until recently, it was thought that the impact of a single rearrangement was critical: either it generated a genomic disorder or unviable cells, or, if it had a selective advantage, it led directly to isolation and speciation. Recently however, an unexpected amount of healthy polymorphic rearrangements have been detected in the human genome [**?**]. For instance,

Tuzun *et al.* hypothesized 297 rearrangements (greater than 8kb) in one single individual relatively to the current genome assembly [**?**]. Finally, other possible causes, or at least factors having a potential influence in the rearrangement process must yet be investigated. This includes, for instance, the spatial arrangement of genomes and what has been called the chromosome territories [**?**, **?**]. Very recently for instance, Branco and Pombo [**?**] suggested a possible role of the intermingling of such territories in translocation.

Human chrom. (reference)	LCR	LCR-mediated	Other features	Hotspot
2 ([?])	+	+		+ (polymorphism)
4 ([?])	-	-	• a deletion of 4 kb in the chimpanzee breakpoint region • more LINE and LTR elements • GC poor regions • an inverted repeat of 5 kb • $(A/T)_n$ rich region	-
5 ([?])	-	-	• 2 direct repetitions (5 pb) • poly-pyrimidine and poly-purine "tracts" (7 & 16 pb) • 2 alternative purine-pyrimidine tracts (10 & 6 pb) • associated to 2 deletions (7 et 9 pb) and one duplication (11 pb) • 2 common fragile sites	+ (chicken, mouse and rat) + (cancer and genomic disorders)
9 ([?])	+	-	-	-
12 ([?])	+	-	duplications appeared simultaneously to the inversion, filling up the DNA sticky ends	-
15 ([?])	+	-	human-specific duplications in breakpoints of a chimpanzee-specific inversion	+ (duplication)
16 ([?])	+	+	• satellites • short direct repetitions • LINE and Alu elements	+ (same inversion in gorilla, tumor)
17 ([?])	-	-	• rich in Alu and LTR elements • a fixation site of topoisomerase II • 1 pentamer of sequence GGGGT	-
18 ([?] et [?])	+	+	Alu elements	-

Table 2: Human-chimpanzee breakpoint regions analysis. Except on chromosome 2 where there is a fusion breakpoint, all are pericentric inversion breakpoints. A + in the hotspot column means that the breakpoint has the same localisation as other rearrangements, either evolutionary (in other species), polymorphic, or involved in inherited disease or cancer.

species and chrom. (reference)	LCR	LCR-mediated	Other features	Hotspot
MMU10-HSA21/22 ([?])	+	-	• mouse-specific repetitions • numerous repetitions on HSA21 • numerous short rearranged blocs on HSA21, flanked by IGL repeats.	-
HSA19-MMU10 ([?])	-	-	rich in simple tandem repeats (such that $(TCTG)_n$, $(CT)_n$ or $(GTCTCT)_n$)	-
Gorilla translocation 4-19 ([?])	+	+	-	+ (Charcot-Marie Tooth disease)
HSA3-MMU (1 bkpt) ([?])	-	-	• sites of deletion in YACs • clusters of same family genes • late replication (similarity with fragile sites) • motif $(TATAGA)_{11}$ which can adopt a hairpin-like secondary structure.	+ (tumor)
HSA7-primates (4 bkpts) ([?])	+	+	-	-
HSA3-primates ([?])	+	?	• GTGG tracks • MIR (mammalian interspersed repeat) • simple repeat and low complexity • retrotransposons • alternative purine-pyrimidine tracks	-

Table 3: Other punctual mammalian breakpoints analysis. A + in the hotspot column means that the breakpoint has the same localisation as other rearrangements, either evolutionary (in other species), polymorphic, or involved in inherited disease or cancer.