Deliverable D2.2
Modeling Evolutionary Events

Deliverable Number	**D2.2**
Deliverable Title	**Modelling Evolutionary Events**
Type of Document	**Report**
Dissemination Level	**Public**
Workpackage	**WP2**
Lead Beneficiary	**University of Malaga**
Contractual Delivery Date	**31.07.2014 (M18)**
Actual Delivery Date	**30.10.2014 (M21)**
Editor(s)	**O.Trelles**
Author(s)	**O.Trelles, J.Karlsson, O.Torreño, J. Arjona, N.Chelbat**
Quality reviewer(s)	**P. Heinzlreiter, A. Upton, J. Perkins**

Document Log

Version	Author(s)	Date	Modifications
V0	O. Trelles, J. Arjona	20.05.2014	Organizational issues
V1	J. Karlsson, O.Trelles	30.05.2014	
V2	J.Arjona, O. Torreño, N.Chelbat	20.06.2014	EE, SFILE and jORCA U/L, Flipper
V3	J.Arjona,	09.10.2014	Sections 2.3, 2.4
V4	N. Chelbat, J. Arjona, J. Karlsson	30.10.2014	General review + formatting and references

MODELING EVOLUTIONARY EVENTS

This section is divided into

1 ABOUT SYNTENY BLOCKS

1.1 INTRODUCTION

Designed methods under classical comparative genome's paradigms that were developed during the 80' become obsolete with the appearance of new high throughput sequencing techniques which allow complete genomes to be sequenced. This fact demands a shift in the sequence comparison paradigms since classical alignment algorithms -prepared for genes and proteins- handle local mutations: insertions, deletions and substitutions. However, other kind of evolutionary events which operate in large-scale DNA -also called genome rearrangements- are not managed by traditional methods. They ignored gen order or similar regions order between sequences and therefore new comparative methods, concepts and definitions at genome level are needed.

In addition, technical improvements in DNA-sequencing techniques have increased data availability as well as the size of data sets. From a computational point of view this represents a challenge for both space and time requirements for computing all this information. The increase in the availability of complete sequenced genomes and the importance of evolutionary processes affecting the organisms' history has increased the interest in studying the underlying molecular evolutionary events (EEs).

The study of EEs between species becomes indispensable when looking for an answer to essential questions regarding the origin of life: *from where we came from, who are we, in which point of evolution are we standing and where are we going.* EEs and evolutionary distances between species based on genome rearrangements have been widely studied by several authors though there is still much to be done. In accordance to this, and based on the fact from evidences that EEs are far from random but fitting some particular distribution, our proposal is to design a new distance measurement that takes into account EEs frequencies.

We are going to describe what synteny block and break point means in the next subsection.

1.2 SYNTENY BLOCKS AND BREAK POINTS

The notion of conserved segments was introduced in 1984 by Nadeau and Taylor in [7]. In the context of comparative genomics, conserved segments are regions where genes content is the same and gene order is conserved. According to this definition, the length of conserved segments could be used for estimating rates of chromosomal evolution. Long

after, Pevzner and Tesler introduced the notion of synteny blocks as regions that can be converted into conserved segments by micro-rearrangements [8]. In addition, they claim that synteny blocks do not necessarily represents areas of continuous similarity, expanding the definition of conserved segments.
When trying to find synteny blocks through alignment methods, one has to face the fact of finding false orthologous regions which makes it harder to define such areas. Thus is necessary to set-up some sort of filtering process in order to leave out those kind of regions.
Sankof et al. designed a method for extracting synteny blocks from comparative maps [1]. They formulated the problem as a Maximum Weight Independent Set (MWIS) search. Their strategy was based on four steps: first, the construction of a set of pre-strips which are of certain length from the common subsequences of each chromosome from each genome; second, the extraction from this set of a subset of mutually compatible (non-intersecting) pre-strips containing a maximum number of markers; third, the addition to this subset of any markers that do not increase the rearrangement distance between the genomes; and fourth, the assembling of the synteny blocks from the markers in the solution. However, this method can build up false synteny blocks by matching pairs of markers as orthologs through an unclear homology alignment between the two genomes. GRIMM-Synteny [10] or OrthoCluster [11] are two examples of the many software methods implemented for extracting synteny blocks.

Lemaitre and Sagot developed a method for breakpoint detection described in their work ”Precise detection of rearrangement breakpoints in mammalian chromosomes" [14]. In the article they try to clarify the origin of the term “*breakpoint*” and the confusion that can be originated from the prefix *break* and from its suffix *point*:

1. *break* suggests a physical break of the sequence but gives an improper biological meaning. A *breakpoint* breaks the conserved segment in the sense that gene order between species sequences is disrupted. For example, if we are comparing human versus mouse genomes and we have ABC segments in the human and ACB segments in the mouse, the region between A and C, which is B appears as a breakpoint.
2. *point* is the second reason why *breakpoint* name can be confusing. A breakpoint is defined as a region between two successive conserved segments. In any case, we can define a breakpoint as a pair of points, Bi (Bstart, Bend) where Bstart(x,y) and Bend(x,y) are the corresponding coordinates in the genomes.

Therefore a breakpoint far from being just a single genomic position or nucleotide, can be of the order of ten to thousands bp length and is always related to synteny blocks.

1.3 REARRANGEMENT EVENTS

During the replication process of DNA, inter- and intra-chromosomal exchanges may vary the order and number of genes in the new chromosome. These exchanges can duplicate genes or even cut them if the new gene is inserted in the middle of another gene. When two genomes from related species are compared, some regions where content and order are preserved are observed. We call these regions conserved segments [1] or ultra-conserved elements [11]. The more divergent or distant the compared species are, the shortest these regions are. In many cases we could extend conserved segments by clustering two or more adjacent genes as well as intergenic sequences into synteny blocks [26]. Between two of those adjacent conserved segments, regions of high probability of chromosome rearrangements or the so called breakpoints can be identified. Thus, the number of breakpoints or the number of conserved segments can be used as a rough measure of their genomic distance [1].

Let $S_1=g_1, g_2, g_3, .., g_n$ be a set of n_{s1} genes in sequence S_1 and $S_2=g'_1, g'_2, g'_3, .., g'_n$ be a set of n_{s2} genes in sequence S_2. The *reverse* of a gene is noted by g_i. A *uni-chromosomal* genome has a single sequence of genes, and a *multi-chromosomal* genome has two or more sequences C_1. Is noteworthy to say that during replication process of DNA, identical copies are not guaranteed due to many factors:

- Mutation: In 1910 T. H. Morgan found out mutations happened spontaneously (T.H Morgan). They could be by replication fault or displacement during replication.
- Recombination: Recombination is the exchange between two homologous sequences of DNA.
- Transpositions: Transposition is a spontaneous process where a DNA sequence is copied or cut and inserted in a new placement in the same genome. The process can be *replicative*, if the sequence is copied and inserted somewhere thus increasing the length of its genome; or *non-replicative* if the sequence is cut and placed somewhere thus keeping the length of its genome.

All these processes lead as a consequence to an amount of different changes or events. For a detailed explanation we use gene representation. Note that some of these events involve sequences that do not need necessarily to be a single gene or set of completed genes if the considered sequence has more than one.

1. A *reversal* (or *inversion*) is the result of change a sequence into its reverse.

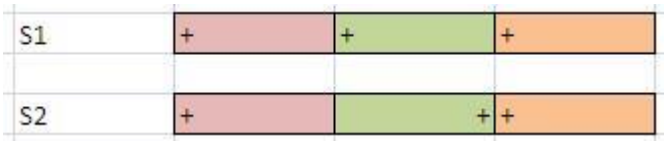

Figure 1. Example of a reversal event

- Before *reversal*: $S_1 = g_1, g_2, g_3$
- After *reversal*: $S'_2 = g'_1, -g'_2, g'_3$

2. A *transposition* (*non-replicative)*. A sequence is cut and placed somewhere in the genome:

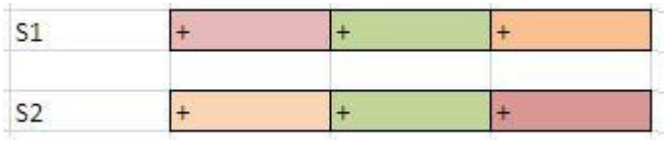

Figure 2. Example of a (non-replicative) transposition event

- Before *transposition*: $S_1 = g_1, g_2, g_3$
- After *transposition*: $S'_2 = g'_3, g'_2, g'_1$

3. A *duplication* (*replicative)*. A sequence is copied and placed somewhere in the genome increasing its length:

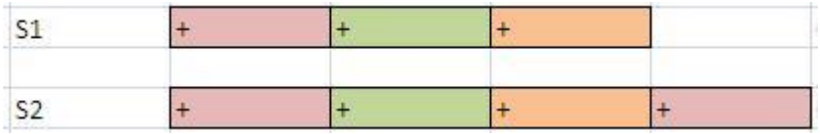

Figure 3. Example of a duplication event

- Before *duplication*: $S_1 = g_1, g_2, g_3$
- After *duplication*: $S'_2 = g'_3, g'_1, g'_2, g'_3$

4. A *translocation* between two chromosomes is the process where two chromosomes exchange their sequences:

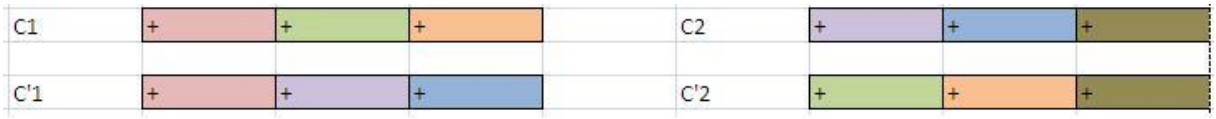

Figure 4. Example of a translocation event

- Before *translocation*: $C_1 = g_1, g_2, g_3$ and $C_2 = g_4, g_5, g_6$
- After *translocation*: $C'_1 = g'_1, g'_4, g'_5$ and $C'_2 = g'_2, g'_3, g'_6$

Two special cases are *fusions*, when one of the two chromosomes turns empty:

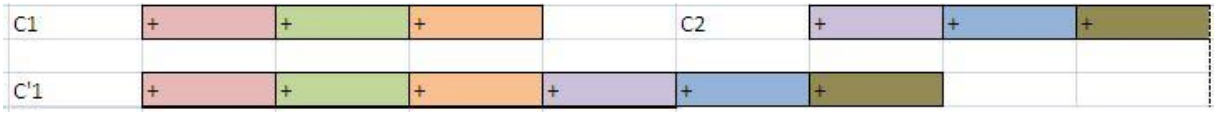

Figure 5. Example of a fusion event

- Before *fusion*: $C_1 = g_1, g_2, g_3$ and $C_2 = g_4, g_5, g_6$
- After *fusion*: $C'_1 = g'_1, g'_2, g'_3, g'_4, g'_5, g'_6$

and *fissions*, when a chromosome is split into two new chromosomes:

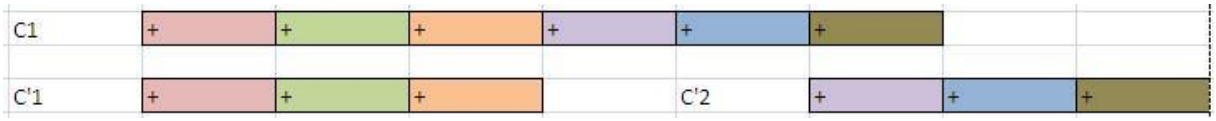

Figure 6. Example of a fission event

- Before *fission*: $C_1 = g_1, g_2, g_3, g_4, g_5 g_6$
- After *fission*: $C'_1 = g'_1, g'_2, g'_3$ and $C'_2 = g'_4, g'_5, g'_6$

1.3.1 Formalism for genome rearrangements and rearrangements processes

Here we describe a formal representation of genome rearrangements and the processes linked to them.

A *gen* can be represented as an integer and a *chromosome* as a set of integers thus, a set of genes. A *genome* is a collection of chromosomes.

When we compare two chromosomes we assume that one of them is the identity permutation I={1,2,3…n}, and the other one is a permutation based on identity chromosome, $\pi = \{\pi_1, \pi_2, \pi_3, \dots, \pi_n\}$.

A reversal operation $r(i,j)$ is the transformation that reverses from i_{th} to j_{th} position:

$$(\pi_1, \dots, \pi_{i-1}, \boldsymbol{\pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j}, \pi_{j+1}, \dots, \pi_n) \rightarrow (\pi_1, \dots, \pi_{i-1}, \boldsymbol{\pi_j, \pi_{j-1}, \dots, \pi_{i+1}, \pi_i}, \pi_{j+1}, \dots, \pi_n)$$

A transposition operation $t(i,j,k)$ is the transformation that 'cuts' the region between i_{th} position and j_{th} position, and paste it in the k_{th} position. In this example $> j > i$.

$$(\pi_1, \dots, \pi_{i-1}, \boldsymbol{\pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j}, \pi_{j+1}, \dots, \pi_{k-1}, \pi_k, \pi_{k+1}, \dots, \pi_n)$$
$$\rightarrow (\pi_1, \dots, \pi_{i-1}, \pi_{j+1}, \dots, \pi_{k-1}, \boldsymbol{\pi_k, \pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j}, \pi_{k+1}, \dots, \pi_n)$$

A reversal transposition $rt(i,j,k)$ is a combined transformation of transposition and reversal operation:

$$(\pi_1, \dots, \pi_{i-1}, \boldsymbol{\pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j}, \pi_{j+1}, \dots, \pi_{k-1}, \pi_k, \pi_{k+1}, \dots, \pi_n)$$
$$\rightarrow (\pi_1, \dots, \pi_{i-1}, \pi_{j+1}, \dots, \pi_{k-1}, \boldsymbol{\pi_k, \pi_j, \pi_{j-1}, \dots, \pi_{i+1}, \pi_i}, \pi_{k+1}, \dots, \pi_n)$$

A translocation operation $T(C_1, i, j, k, C_2)$ over a chromosome $C_1 and\ C_2$ is the transformation that 'cuts' the region i_{th}, j_{th} and pastes in k_{th}position in C_2 chromosome, where $C_1 \neq C_2$.

$$C1(\pi_1, \dots, \pi_{i-1}, \boldsymbol{\pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j}, \pi_{j+1}, \dots, \pi_{k-1}, \pi_k, \pi_{k+1}, \dots, \pi_n)$$
$$C2(\pi_1, \dots, \pi_{i-1}, \pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j, \pi_{j+1}, \dots, \pi_{k-1}, \boldsymbol{\pi_k}, \pi_{k+1}, \dots, \pi_n)$$
$$\rightarrow C1(\pi_1, \dots, \pi_{i-1}, \pi_{j+1}, \dots, \pi_{k-1}, \pi_k, \pi_{k+1}, \dots, \pi_n)$$
$$\rightarrow C2(\pi_1, \dots, \pi_{i-1}, \pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j, \pi_{j+1}, \dots, \pi_{k-1}, \boldsymbol{\pi_k, \pi_i, \pi_{i+1}, \dots, \pi_{j-1}, \pi_j}, \pi_{k+1}, \dots, \pi_n)$$

A fusion operation $Fu(C_1, C_2)$ over a chromosome $C_1 and\ C_2$ is the transformation that 'joins' two chromosomes into one.

$$C1(\pi_1, \pi_2, \dots, \pi_n)$$
$$C2(\pi_1, \pi_2, \dots, \pi_n)$$
$$\rightarrow C1(\pi_1, \pi_2, \dots, \pi_n, \pi_1, \pi_2, \dots, \pi_n)$$

A fission operation $Fi(C_1, i)$ over a chromosome C_1 is the transformation that 'splits' one chromosome into two new chromosomes.

$$C1(\pi_1, \dots, \pi_{i-1}, \boldsymbol{\pi_i}, \pi_{i+1}, \dots, \pi_n)$$
$$\rightarrow C1(\pi_1, \dots, \pi_{i-1}, \boldsymbol{\pi_i})$$
$$\rightarrow C2(\pi_{i+1}, \dots, \pi_n)$$

1.4 SORTING PERMUTATIONS

Methods to formulate chromosomal rearrangements have been developed. However the most known ones were developed in the last 20 years under the term sorting permutations methods. The first strike started in 1993 by Kececioglu and Sankoff [1].They proposed both, an approximation algorithm and an exact algorithm to sort unsigned permutations by reversals, in $O(m(L(n,n))$ time and $O(n^2)$ space, where $L(n,n)$ is the time to solve a linear programming of n variables and n constrains, and m is the size of the branch-and-bound search tree. Bafna and Pevner improved the method for signed and unsigned permutations [2][3]. In [28] Hannenhalli and Pevzner presented the first polynomial time algorithm for sorting signed permutations by reversals. Caprara in [27] demonstrated that sorting by reversals is a NP-hard problem.

Bafna and Pevzner were the first on studying transposition permutations [3][8]. The problem was defined as: given two permutations, the sorting by transpositions is to find a shortest

series of transpositions to transform one permutation into the other. The first approximation algorithm had $O(n^2)$ time complexity.

In practice, scientists have observed that transpositions and reversals occur with different frequency [4]. Hannenhalli and Pevzner [27], Eriksen [13], Dias and Meidanisc[29], are studying this problem. The double cut and join operation [6] allows an efficient sorting dealing with translocations, inversions, fissions, fusion.

2 FEATURING BREAKPOINTS

2.1 INTRODUCTION

There are chromosomal rearrangements events, well known that define the evolutionary history of the considered genome and hence the organism it represents, these rearrangements leave some footprints one can follow to find out what did happen during evolution. One such footprint is the so called break points, genomic places where these chromosomal rearrangements are more likely to occur.

As in general breaking points are flanked by constant genomic regions the idea is to find common patterns of these areas of rearrangements across the whole genome. The main task is to apply the approach on patterns for probes classification, but for finding genetic features or “words” that lead us to predict genomic locations for events as Inversion, Translocation, Transposition, Fusion, Fission, Deletion, Segment Duplication, Fragile sites, etc and in a subsequent step pre-visualize Evolution and (maybe) foresee speciation. In this sense some of my raised questions were “why is more likely that these genomic changes do occur in certain regions? Is there any pattern(s) within or out of these breaking points that can be detected and be used to fish existent but unknown breaking points? Can we also use them to predict genomic places with evolutionary impact and therefore give us some hint on next speciation steps? Are these patterns dependent on specific genomic locations regions, functions, specie? Or depends on a combination of some of all these factors?

Considering Synteny blocks as conservative regions in opposition to Breaking point regions, the idea relies on detecting patterns that could differentiate conservative from non-conservative genomic regions within the whole genome. The same formula as the one used for probes sequence classification is used here but for groups of much larger sequences containing demonstrated (orthologous) conservative and non-conservative genomic regions.

Our first attempt was to find such genomic signatures following the work-flow used for the classification of probes as follows:

(i) we download the corresponding genomes (in a first instance complex genomes from mammals) from a public database;

(ii) based on publicly available break points coordinates retrieve the sequences representing break points (coordinates from Sagot 2007-2008) and sequences known to belong to complete conserved regions representing synteny blocks (ensembl);

(iii) clean up the retrieved sequences and prepare the dataset to run under the classification machine (Kernel based SVM);

(iv) we check the resulting better classifiers, thus the best features that can be used to distinguish between one group (break point) and another (synteny blocks). This approach can be directly linked to provide genetic sequences to characterize features to distinguish between recombination 'hot' & 'cold' spots and to explore how do they vary across the genome.

2.2 SUPPORT VECTOR MACHINES

Support Vector Machines are supervised learning models (machine learning task of inferring a function from labeled training-test data) with associated learning algorithms that analyze data and recognize patterns. This SVMs are used for classification such that given a set of training examples, each marked as belonging to one of two categories (in our case belonging to Break points-Non Break points), a SVM training algorithm builds a model that assigns each item of each category or class as a point in a space. The two classes in such space are separated by a gap such that the svm "learns" to separate the items belonging to each class by assigning each point to each side of the gap. The svm then "predicts" where new items belong to a class or another based on which side of the gap they fall on.

Kernel Trick: Used through smv to perform a non-linear classification. By default a svm, maps each item as a point into a two dimensional feature space. By using the kernel trick, items can be mapped into a high-dimensional feature space. Our approach is based on string kernels as we use as items sequence of characters {ACGT}. Two types of kernels will be used, linear and quadratic. Performance of classification results will be measured according to a contingency or confusion matrix where the total numbers of true positives and true negatives will be considered in reference to the total items classified (total number of true and false positives plus total number of true and false negatives). The idea behind it is simply to estimate how good the labeled items are classified in each corresponding class or are classified in "confused" classes.

About **balance-unbalanced dataset** and **distribution length**: a balanced datasets refers to a dataset with equal number of items for each class, whereas in an unbalanced datasets the number of items for each class is different. The distribution of sequence length is considered here when running the linear kernel as the expected performance can be affected. If we run the normalized kernel, we do not actually need an equal length distribution but instead a pool of equal items from positive and negative class. That means if we have ranked all sequences length then we can select those sequences on the very top of the ranking that are corresponding for both positive and negative class to be the longer ones.

kernel methods *are a class of algorithms for pattern analysis, whose best known member is the support vector machine (SVM). The general task of pattern analysis is to find and study general types of relations (for example clusters, rankings, principal components, correlations, classifications) in datasets. For many of these tasks, data have to be represented as feature vectors, but kernel methods replace this representation by similarities to other data points. Any linear model can be turned into a non-linear model by applying the "kernel trick" to the model: replacing its features (predictors) by a kernel function.*

String kernel is a kernel function that operates on strings, i.e. finite sequences of symbols that need not be of the same length. String kernels can be intuitively understood as functions measuring the similarity of pairs of strings: the more similar two strings a and b are, the higher the value of a string kernel K(a, b) will be.

		Condition positive	Condition negative	
Test outcome	Test outcome positive	**True positive**	**False positive**	**Precision =** Σ True positive / Σ Test outcome positive
	Test outcome negative	**False negative**	**True negative**	**Negative predictive value** = Σ True negative / Σ Test outcome negative
		Sensitivity= Σ True positive / Σ Condition positive	**Specificity=** Σ True negative / Σ Condition negative	**Accuracy =** Σ True positive + Σ True negative / Σ Total population

Table 1. Components in a Confusion table and how to compute Sensitivity, Specificity, Precision and the Accuracy of the classification.

2.3 STATE OF THE ART

There is not a clear definition of what breakpoint and/or synteny blocks refers to, most authors define breakpoint according to flanking synteny blocks regions and use pairwise comparison between genomes to refine synteny. Under Sagot et all [14], conserved segments are defined as regions found in 2 genomes in which homologous genes retain the same order and relative map position in both genomes. According to this definition, there are approximately 25000 mammalian genes with around 3x10E9 bp meaning a conserved region with 3 genes to be 9,000 000 000 (considering the size of the genome).

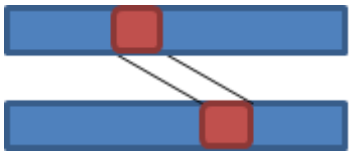

Figure 7. Esquematic representation of a synteny block between two genomes F and G

When comparing 2 genomes Fi and Gi are considered to be as the best reciprocal alignments. They are considered as sequence Anchors or landmarks. Could those sequences be considered as the initial seeds from where synteny blocks can be built? Some authors consider conserved segment according to conserved genes and synteny blocks according to anchors as a set of anchors that appear together BUT not necessarily in the same order [3].

Now, when considering the genomes of 2 different organisms, there are some properties that can be used to describe genomes, namely:

1. Compositional measures like k-word content
2. Fraction of genome represented by transposable elements as tracers for defining the evolutionary histories of groups of alike organisms
3. Sequence organization
4. Identification and characterization of genes
5. Proteomic comparison

All of them are considering within the comparative genomic landscape. Comparative genomics can be understood as comparison within genomes: genome of organism's variation in base composition, k-tuple frequency, gene density, number and types of transposable elements and in identifying any duplicated region. Also comparative genomics can include the comparison between genomes: for those closely related organisms by identifying conserved genes, gene organization and control elements. For distant related organisms by identification of genes restricted to particular clades of Phylogenetic tree [16]. According to this DNA sequences can be modified by

- Substitution (Point mutation)
- Insertion/Deletion: Indels

→ Alignment of regions <= coding genes

- Segmental duplication
- Inversion
- Transposable element insertion
- Translocation

→ Larger than the coding regions of genes meaning possible Breakpoints

All those are evolutionary forces affecting genome's architecture. Larkin and col.[17] studied such effects by comparing synteny relationships among 10 amniotes: Human, chimp, macaque, rat, mouse, pig, cattle, dog, opossum and chicken and concluded that chromosome breakage during evolution is NOT RANDOM.

Evolution acts differently in breakpoints and synteny blocks being breakpoints used to generate new genetic variation and novel combination of genes and regulatory elements for adaptive phenotypes.

Once again pairwise comparison is used as main method to distinguish between 2233 homology synteny blocks and 1064 evolutionary breaking regions for all 10 genomes.

The idea that evolution acts differently is highlighted here by msHSB (multispecie homology synteny blocks) found to be specially enriched for genes associated to development of the central nervous system while msEBR (multispecies evolutionary break regions) are enriched for genes associated with adaptive functions. These regions are also enriched for structural variations like segmental duplications, CNV, indels, retrotransposons, zinc finger genes and SNP.

They define a HSB as a minimum of 2 adjacent markers in the same chromosome in the 2 compared species without interruptions (based on Murphy et al description).

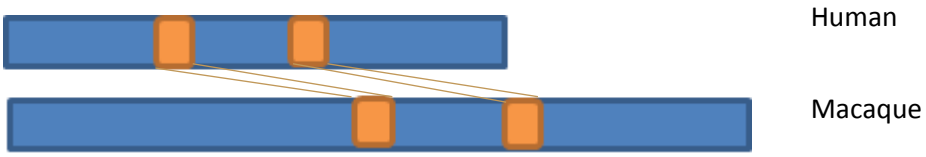

Figure 8. Schematic representation of HSB between humans and macaques

According to this method the inputs and outputs are

- Orthologous gene pairs from whole genome sequence: coordinates from Ensembl BioMart Database v38 with a size of HSB > 500 Kb
- Radiation hybrid map

Algorithm → Chromosomal position Coordinates of HSB in the reference and in the target genomes

The most representative genome signature found in msEBR is segmental duplication [18]. Breakpoints and regions of segmental duplication or low copy repeats (length > 1Kb with >90% identity) are shown recently to co-localise.

The plausible hypothesis is that they co-localise because duplications caused rearrangements (i) or because these are regions with an inherent fragility (ii) consider the NAHR (Non Allelic Homologous Recombination). Duplications appeared as a consequence of the rearrangement as a fill-in of the gaps.

On the contrary, other authors as Sagot consider the IBM (Intergenic Breakage Model) where breakage occurs in a uniform manner across the whole genome and try to demonstrate that within the genome there are regions more susceptible of breaking than others. They leave the evolutionary pressure effect and focus at a heterogeneous breaking point distribution represented by a difference in gene type and number and GC content (higher for BP) across the whole genome (for humans and 5 eutherians).

For both approaches breakpoints are under- represented in genes though for Sagot regions susceptible of breaking are heterogeneously distributed and more likely to be highly transcribed and replicated.

Evolutionary pressure underlies such conclusion as these breaking regions would correspond to less conserved regions and regions susceptible to produce new features for adaptation.

There are studies based on physically labeled synteny blocks whose markers are located by radiation hybrid as described in S.Moore and col [19] being this information more reliable as it is based on physical location rather than on global pairwise or multiple comparison. In their work they define *K markers* contained within *anchors* or pair of orthologous and consecutive genes. And Gk as the directed graph with anchors as vertices, considering conflict types I and II (anchors not appearing in the same orientation in the considered anchors – sharing markers that appear in the middle of two anchors in one genome and as main component in the other genome).

A synteny block is then defined as a subgraph Hk of Gk with no conflict arcs and containing connected blocks with the extremes of genomic coordinates. This leaves out all rearrangements due to translocation and inversions or insertions. Their solution therefore is quite extreme as they do not explain such phenomena avoiding them by exclusion.
Opposite to the literature, they assume just one parameter K as the maximal distance between 2 blocks to be equal to the minimal size of a block to be retained, thus *parameters d* and *s* in just one (*k→ d= s).*

In practice they did exclude centromers: from Ensembl they construct their breakpoints and synteny blocks on Human genome Assembly 35-2004, Mouse assembly 35-2005, and dog CamFam-2005. Then one-to –one orthologous genes as anchors to build the 2 blocks that SHOULD BE CONSECUTIVE in the same chromosome. Then breakpoint refinement and included orthologous sequences at the extremities of the blocks in the aligned sequences.

This approach still yielded break points larger than the ones obtained with the refinement methods of Sagot.

Table 3: Comparison of the distributions of breakpoint sizes between the four datasets.

length	min	max	median	mean
REFINED	21	2,185,434	51,136	128,644
GRIMM2	313	5,418,383	155,816	364,199
GRIMM3	2,490	4,953,520	267,609	454,490
ENSEMBL	2	82,331,123	106,534	1,513,770

Table 2. Comparison of the BP sizes distribution through a Wilcoxon rank sum test

Pevzner [2] used a pairwise comparison method between the released genomes of human and mouse and found out 245 rearrangements that could explain how the 281 synteny

blocks present in the 2 genomes got into their current location. According to their method human and mouse come from the same ancestral but mouse genome had to get into the following changes

- 149 inversions
- 93 translocations
- 3 fissions

If we perform all these changes in the mouse genome then we will end up having the synteny blocks ordered as in the human genome

The same approach was extended to reconstruction between Human-Rat and Mouse genomes through a multiple alignment method [21].

Their method is based considering the first original synteny problem by Nadeau and Taylor [7]. It was observed that the rate of rearrangements in Rat-Mouse is higher than in human but there are rearrangements in all three species considered as hot-spots.

The group of Pevzner tried to find an approach to build up the putative genomic architecture of an ancestral mammalian genome to solve the problem of the genomic distance.
Several authors previously had different approaches as a polynomial-time algorithm that computes a scenario to transform one genome into another through reversals, translocations, fusions and chromosomes.

This approach did yield an estimate number of rearrangements as

- *Microrearrangements fissions* of: intrachromosomal rearrangements small spam
- *Macrorearrangements*: intrachromosomal rearrangements larger spam plus interchromosomal rearrangemens

The synteny blocks generation follows through GRIMM Synteny algorithm which considers 2 important features

- Preservation of microrrearragements information within synteny blocks so the individual synteny history can be studied
- Extendable from 2-3 genomes synteny blocks to study

In summary, GRIMMS algorithm allows the study of micro and macro rearrangements separately and estimate the number of both in all considered genomes, in this case humans, rats and mice finding a final number of

- 417 SB > 299 Kb between Human -Rat
- 394 SB 293 Kb between Human-Mouse
- 162 SB 100 Kb between Mouse-Rat

Through the Multirearrangement algorithm

When considering the 3 ways block algorithm then

- 1533 microrearrangements between H-R
- 1070 microrearrangements between H-M
- 1260 microrearrangements between M-R

Datasets used for comparison	
Human NCBI Build 33	April 2003
Mouse NCBI Build 30	February 2003
***Rat Baylor HGSV v 3.1**	June 2003

Table 3. Datsets used by Pevzner for the multiple alignment comparison method. (*First by using Repeat Masker and then Tandem Repeat Finder

Authors	Method Basis	Used data	Break point definition	SB definition	Number of breakpoints	Whole Human genome coverage
Sagot2008	*Breakpoints number changes depending on the size (300 kb and for 1 Mb) of the anchor used in the two-way (2 species at a time) anchors. Segmentation algorithm plus a refinement of the aligned regions Alignment of each BP sequence from genome 1 against its specific orthologous in genome2	Human NCBI Build 35 HG15 Mouse NCBI Build 35 Dog CamFam 2005	A region between 2 synteny blocks that is consecutive on the reference genome but not in query genome	Unbroken chain of markers which appear in the same order and same orientation in both genomes	H-M: 355 H-D : 240	86.7% (before refinement)
Pavzner 2003	Pairwise comparison through a Multiple Genome Rearrangement Algorithm	Released data from Watertson et all.2002	GRIMM-synteny algorithm plus a two-way-block	Set of non-overlapping two-anchors that can be represented as diagonals in a genomic 2D dot-plot (clusters of larger points)	H-M:246	89.6%
Pavzner 2004	Multiple comparison through a Multiple Genome Rearrangement Algorithm	Human NCBI 33 Mouse NCBI 30 Rat Baylor HGSC v3.1	GRIMM-synteny algorithm plus a three way-block	Set of non-overlapping three-anchors	*H-M: 265/193 H-R: 254/190 M-R:77/56	89.5%
Ensembl V34 [10]	Blast+Repeat mask in whole genome discarding blocks with a min size *min-length*	Human NCBI Build 35	Inter-and intra- chromosomal rearrangement as a structural variation	Regions where both sequence and gene order is conserved between two (closely related) species	H-M: 200	76.8%
M.Larkin (2009)	Pairwise comparison between all 10 amniote genomes	Human NCBI Build 37	Sequence features from UCSC and Bonferroni correction for	Sets of ordered anchor points once	1064 (EBRs)	

			multiple comparison	orthologous gene pairs for all genomes were downloaded and cattle and pig RH-fingerprint maps were used to map common regions		

Table 4. Summary of some of the most known methods to feature synteny blocks and break points.
*Breakpoints considered as of re-use regions, changes in number depending on the size (300 kb and for 1 Mb) of the anchor used in the two-way (2 species at a time) anchors.

2.4 METHODS

2.4.1 REQUIRED SOFTWARE-LIBRARIES-GENOMES

The required libraries should be installed on the perl path or explicitly redirects the script we run to the path where the libraries needed are located.

Libraries: We need from the BioInfII library the Preprocessing.pm and from and Proc the Background.pm script. Make sure the library is in the right path: go to CPAN and install Proc:Backrgound. In case you have the right as administrator to install it in perl libraries, fine. If not we can create our own perl libraries and redirect the path there.
In this case we do not need to use Alignment.pm BUT preprocessing.pm because we will adapt the strings from the sequences to have numerical values and so be "understood" by libSVM.

BP2Fasta: Need to have it installed in order to retrieve sequences from genome coordinates. Any other program or way of retrieving genomic sequences from coordinates can also be used (i.e Extract Genomic DNA- in Galaxy, or getSeq function from BioString).

Required Genomes: Need to have downloaded and stored the genomes to be used as reference. We will retrieve the sequences based on coordinates using such genomes (or chromosomes). When downloading a reference genome, make sure the version corresponds to the version from where the coordinates of your input sequences were taken. In our case some coordinates for breakpoints were taken from an old version of the human, mouse and dog genomes (NCBI human assembly Gr35), thus we downloaded the same version to retrieve the corresponding sequences. We will also the last released human genome assembly (NCBI human assembly Gr38) to generate the negative class to be used in our classification tasks.

2.4.2 DATASETS GENERATION

We used publicly available data to perform our classifications. The idea is to have sequences representing the two classes for the classification task, thus a positive and a negative class.

We consider as positive class the one containing breakpoints sequences and as negative class the one containing Known conservative genomic regions, well known to be "No Breakpoints"; in this case sequences from synteny blocks.

Each dataset is composed of a pool of sequences from both breakpoints and synteny blocks.
There is not a universal-unique definition of a breakpoint; therefore we will consider datasets based on freely available co-ordinates and datasets based on self-generated co-ordinates. The latest will be the ones to support our potential definition of breakpoints as regions upstream-downstream or in the middle of 2 consecutive homology regions. Datasets will cover not just prokaryote genomes as bacteria but also genomes from primates and other mammals. For all co-ordinates we have to retrieve the corresponding sequences that will be used as inputs in our classification task. In order to get such sequences we had .txt files containing headers as start and end of each sequence.

3. BIBLIOGRAPHY

[1] Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement, J. Kececioglu and D. Sankoff, *Algorithmica*, 1995.

[2] Genome rearrangements and sorting by reversals, V. Bafna and P. Pevzner, *SIAM J. Comput.*, vol. 25, no. 2, pp. 272–289, 1996.

[3] Sorting by Transpositions, V. Bafna and P. a. Pevzner, *SIAM J. Discret. Math.*, vol. 11, no. 2, pp. 224–240, May 1998.

[4] "Parametric genome rearrangement.," M. Blanchette, T. Kunisawa, and D. Sankoff ,*Gene*, vol. 172, no. 1, pp. GC11–7, Jun. 1996.

[5] (1 +) -Approximation of sorting by reversals and transpositions, N. Eriksen, vol. 289, pp. 517–529, 2002.

[6] Efficient sorting of genomic permutations by translocation, inversion and block interchange. S. Yancopoulos, O. Attie, and R. Friedberg, *Bioinformatics*, vol. 21, no. 16, pp. 3340–6, Aug. 2005.

[7] Lengths of chromosomal segments conserved since divergence of man and mouse. J. H. Nadeau and B. A. Taylor, Proceedings of The National Academy of Sciences, 81:814–818, 1984.

[8] Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes. Pavel Pevzner and Glenn Tesler, Genome Research, 13:37–45, 2003.

[9] Algorithms for the Extraction of Synteny Blocks from Comparative Maps. Vicky Choi, Chunfang Zheng, Qian Zhu, and David Sankoff. 2007.

[10] GRIMM: genome rearrangements web server. Glenn Tesler. Bioinformatics/computer Applications in The Biosciences, 18:492–493, 2002.

[11] OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics. Xinghuo Zeng, Matthew J. Nesbitt, Jian Pei, Ke Wang, Ismael A. Vergara, and Nansheng Chen ,In Extending Database Technology, pages 656–667, 2008.

[12] Precise detection of rearrangement breakpoints in mammalian. Claire Lemaitre , Lamia Zaghloul , Marie-France Sagot Christian Gautier , Alain Arneodo ic Tannier and Benjamin Audit. BMC Bioinformatics 2008,9:286

[13] Eriksen, N., 2002. (1 +) -Approximation of sorting by reversals and transpositions. , 289, pp.517–529.

[14] Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation.. Claire Lemaitre , Lamia Zaghloul , Marie-France Sagot Christian Gautier , Alain Arneodo ic Tannier and Benjamin Audit. BMC Genomics 2009, 10:335

[15] Genomic features in the breakpoint regions between synteny blocks. Phil Trinh, Aoife McLysaght and David Sankoff. Bioinformatics, 2004

[16] Computaional Genome Analysis. An introduction. Richard C.Deonier, Simon Tavaré and Michael S.Waterman. Springer 2005

[17] Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Denis M. Larkin, Greg Pape, Ravikiran Donthu, et al. Genome Research, 2009 19.

[18] Segmental duplications: an 'expanding' role in genomic instability and disease. Beverly S. Emanuel & Tamim H. Shaikh. Nature Reviews Genet4cs 2, 791-800 (October 2001)

[19] High resolution radiation hybrid maps of bovine chromosome 19 and 29: comparing with the bovine genome sequence assembly. Parna Prasad, Thomas Schiex, Stephanie McKay, Brenda Murdoch, Zhiquan Wang, James E Womack, Paul Stothard and Stephen S.Moore. BMC Genomics, 2007, 8: 310

[20] Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes. Pavel Pevzner and Glenn Tesler. Genome Research. 2003 13.

[21] Reconstructing the Genomic Architecture of Ancestral Mammals: Lessons From Human,Mouse, and Rat Genomes. Guillaume Bourque, Pavel A. Pevzner, and Glenn Tesler. Genome Research, 2004

[22] Lengths of chromosomal segments conserved since divergence of man and mouse. J. Nadeau and B. Taylor. Proceedings of the National Academy of Sciences USA, 81:814–818, 1984

[23] Ensembl 2014. Nucleic Acids Research 2014 42 Database issue:D749-D755 http://www.ensembl.org/info/genome/variation/data_description.html#classes

[24] Sequencing human-gibbon breakpoints of synteny reveals mosaic new insertions at rearrangemet sites. Sanhosh Girijaran, Lin Chen, Tina Graves et al. Genome Research, 2009 19

[25] A fine-scale map of recombination rates and hotspots across the human genome. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. Science 2005, 310:321-324

[26] A Flexible Ancestral Genome Reconstruction Method based on Gapped Adjacencies. Y. Gagnon, M. Blanchette and N. El-Mabrouk. BMC Bioinformatics, 13 (Suppl 19): S4, 2012

[27] Sorting by reversals is difficult. Proceedings of the first annual international conference. Caprara, A., 1997 Available at: http://dl.acm.org/citation.cfm?id=267531 [Accessed July 6, 2014].

[28] To Cut... or Not to Cut (Applications of Comparative Physical Maps in Molecular Evolution). Hannenhalli, S. & Pevzner, P., 1996, SODA, pp.304–313.

[29] Genome rearrangements distance by fusion, fission, and transposition is easy. Dias, Z. & Meidanis, J., 2001. Proceedings Eighth Symposium on String Processing and Information Retrieval, pp.250–253. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=989776.