The Potential of Family-Free Genome Comparison

Marília D. V. Braga[1], Cedric Chauve[2,3], Daniel Doerr[4,5], Katharina Jahn[4,5], Jens Stoye[4,5,⋆], Annelyse Thévenin[4,5], and Roland Wittler[4,5]

[1] Inmetro, Duque de Caxias, Brazil
[2] Department of Mathematics, Simon Fraser University, Burnaby BC, Canada
[3] LaBRI, Université Bordeaux I, Talence, France
[4] Genome Informatics, Faculty of Technology, Bielefeld University, Germany
[5] Institute for Bioinformatics, CeBiTec, Bielefeld University, Germany

Abstract. Many methods in computational comparative genomics require gene family assignments as a prerequisite. While the biological concept of gene families is well established, their computational prediction remains unreliable. This paper continues a new line of research in which family assignments are not presumed. We study the potential of several family-free approaches in detecting conserved structures, genome rearrangements and in reconstructing ancestral gene orders.

1 Introduction

In more than 20 years of research in computational comparative genomics [44,48] a large variety of questions have been addressed. By now, strong methods are available to study the structural organization of genomes as well as to unravel their shared and individual evolutionary histories. The structural organization of genomes does not only give insights into species' phylogeny, but also hints at interactions within and between sets of genes by means of their involvement in metabolic and regulatory networks. As such, one aims to understand cell functions. Whereas point mutations generally affect one or a few nucleotides, large-scale mutations such as rearrangements, deletions, substitutions, or insertions affect one or more genes. These modifications alter the structural organization of the genome which can cause profound changes in the cellular machinery. Identifying and quantifying such structural modifications is crucial in understanding the highly complex functions of organisms and their interactions with the natural environment.

Initial approaches to study genome rearrangement considered pairwise comparisons with well identified one-to-one orthologous markers [44], for many of which polynomial time algorithms for computing distances and evolutionary scenarios could be designed [5, 6, 30, 46, 63]. Extensions considering more than two genomes lead to hard problems [8, 13, 15, 41, 46, 61], with few exceptions [26, 54]. David Sankoff initiated formulations and algorithms for genome rearrangement

⋆ corresponding author, jens.stoye@uni-bielefeld.de

problems with duplicated markers originating from gene families [45], quickly followed by the outline of a general approach that would consider both gene orders and gene family information as input to genome rearrangement problems [49]. Since then, genome rearrangement with unequal gene content and gene families, where genomes are represented by signed sequences, has been intensively explored; for reviews see [16, 27].

Another line of research in computational genomics aims at the detection of genomic segments that are conserved across different species. The presence of such structures often hints at functional coupling of the contained genes, or indicates remnant ancestral gene order which is valuable information for phylogenetic reconstruction. Initial approaches in this field — like early rearrangement studies — required the identification of one-to-one orthologous markers [4,32,33], but in the following most of them were adapted to a more general genome model that allows genomes to differ in their marker set and to have homologous markers on the same genome [21, 31, 50].

All of the above methods, that we call *family-based*, require prior gene family assignments. However, biological gene families are difficult to assess; commonly, they are predicted computationally. In doing so, they can be either obtained from databases [42, 55, 59] or directly computed based on the particular dataset under consideration [36, 40, 51]. In either case, the obtained assignments are predicted by some computational method which typically involves a clustering phase in which genes are partitioned into groups representing the predicted families. Generally, the results of such efforts depend on arbitrary parameters of sequence comparison, similarity quantification and clustering. These parameters are user-controlled and influence the size and granularity of the computed gene families. In particular, when genes within biological gene families are largely diverged, computational means may not be able to resolve gene family assignments accurately [28]. Consequently, errors are introduced into the primary dataset which deteriorate subsequent analyses, a phenomenon that can be amplified when phylogenetic trees for the gene families are considered [16, 39]. The quest to reduce misassignments in gene family construction also led to the use of positional homology [9, 57, 58, 65].

Recently, in an attempt to avoid these problems, a *family-free* method, that does not assume prior gene family assignment, has been proposed for computing the adjacency score between two genomes [22]. In this approach, given the gene similarities, the aim is to find pairwise gene assignments while maximizing the conserved adjacency measure. In other words, next to finding the maximal number of adjacent genes along different genomes, the method also infers homologies between genes. It should be noted that these homologies are not equivalent to gene families in the classical sense, as by design only one-to-one relationships are detected, while a gene family in general may consist of a potentially large set of orthologous and paralogous genes. Given the nature of the detected one-to-one relationships, they are not unlikely to form sub-families of biological gene families. Therefore they can be further utilized in gene family construction.

Here we go beyond this one application and explore how various problems in computational comparative genomics could be approached in a family-free setting. We do not necessarily provide full solutions to the proposed problems.

This paper is organized as follows. After basic definitions in Section 2, we extend earlier results on the adjacency measure to more than two genomes and to larger conserved structures (gene clusters) in Section 3. A more dynamic view is taken in Section 4, where we apply the ideas to rearrangement distances, most notably the Double Cut and Join distance. In Section 5, finally, we indicate how the family-free approach could be further extended to the reconstruction of ancestral genomes. The paper concludes with a discussion in Section 6.

2 Basic Definitions

A chromosome is a DNA molecule composed of antiparallel strands and can be read in either of the two possible directions. Since each gene, representing an interval along the DNA, lies in one of the two strands of the chromosome, the orientation of the gene depends on the adopted reading direction. The representation of a gene g in a chromosome can then be the symbol g, if it is read in direct orientation, or the symbol $\overline{g}$, if it is read in reverse orientation. Without loss of generality, we will assume in this paper that each chromosome has a canonical reading direction, giving a natural left to right order of its genes.

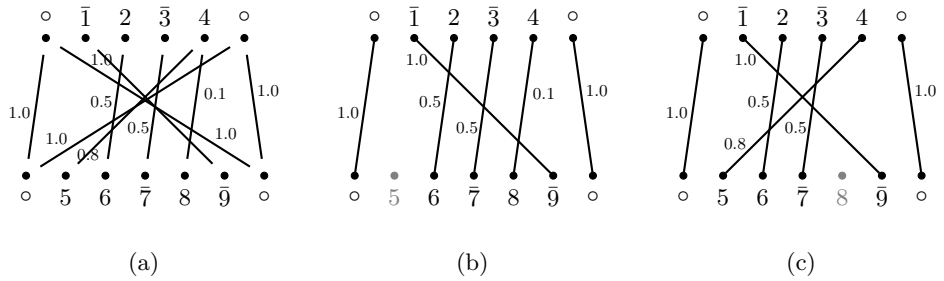
A genome consists of one or more chromosomes that can be either linear or circular. For ease of presentation, throughout this paper we will consider only unichromosomal linear genomes. The general case can be easily inferred with minor modifications.

A unichromosomal linear genome is represented as a sequence of distinct symbols, flanked by telomeric ends indicated by the $\circ$ sign: $G = (\circ\ g_1\ g_2\ \ldots\ g_n\ \circ)$. The size of G with n genes and two telomeric ends is $|G| = n + 2$. When we consider a set of genomes, we will assume that all genes can be distinguished from each other, i.e., every two genomes $G \neq H$ share only the telomeric ends.

Let $\mathcal{A}$ be the universe of all genes and let $\sigma : \mathcal{A} \times \mathcal{A} \rightarrow [0, 1]$ be a *normalized similarity measure* between all pairs of genes.

Definition 1 (Gene similarity graph). *For a set of k genomes $\{G_1, \ldots, G_k\}$, the* gene similarity graph *is defined as an ordered weighted undirected k-partite graph $B = (G_1, \ldots, G_k, E)$, where each gene and each telomere represents a node, and the nodes are ordered following the chromosomal order. Any two genes g and h, belonging to two distinct genomes, are connected by an edge $e_{g,h} \equiv \{g, h\} \in E$ with weight $w(e_{g,h}) := \sigma(g, h)$, if and only if $\sigma(g, h) > 0$. Telomeres in distinct genomes are always connected with edges of weight* 1.

We call a gene $g \in G$ *unconnected* if there exists no other gene h in any of the other genomes $H \neq G$ such that $\sigma(g, h) > 0$. An example of a gene similarity graph for the case $k = 2$ is shown in Figure 1(a). The k-partite gene similarity graph features similarity relationships between genes of different genomes whereas similarities between genes within the same genome are ignored.



(a) (b) (c)

Fig. 1. (a) Example of a gene similarity graph for $k = 2$. Part **(b)** shows a matching in which the weak edge with weight 0.1 between genes 4 and 8 is selected, creating a conserved adjacency between $(\bar{3}, 4)$ and $(\bar{7}, 8)$. In the matching of **(c)** the stronger edge with edge weight 0.8 between genes 4 and 5 is selected.

For now, if information about paralogous relationships between genes within the same genome is desired, it must be gained through a postprocessing step incorporating the results obtained by the methods presented herein.

3 Detecting Conserved Structures

Many gene order studies quantify conserved structures based on well-defined proximity relations between the chromosomal locations of pairs or groups of genes. Typical proximity relations between pairs of genes are conserved adjacencies [44,47,60] and generalized conserved adjacencies [62], whereas proximity relations between groups of genes include common intervals [21,33,50,56], max gap clusters (gene teams) [4,31], approximate common intervals [12,34,43], generalized adjacency clusters [64,67], and conserved intervals [6]. We discuss conserved adjacencies in Section 3.1 and common intervals and some of its derivatives in Section 3.2.

Whenever one-to-one relationships between genetic markers, genes or genome segments (identified through some proximity relation) between genomes must be established, comparative genomics applications commonly incorporate matchings. For example, in aligning whole genomes, one aims to find a matching between genome segments that maximizes the similarity of the respective sequences, but also minimizes the number of breakpoints (or other measures of structural dissimilarity) in the final ordering of segments [19]. Similarly, recent methods in predicting co-orthologs and gene families not only assess the sequence similarity between genes, but also their position within the genome [20]. In the following we describe approaches that incorporate matchings to identify conserved adjacencies and common intervals without the use of gene family assignments.

3.1 Conserved Adjacencies

Previous work. Two genes that are located next to each other in a genome are said to be adjacent, their adjoining extremities form an *adjacency*. An early measure for family-based genome similarity was to count the number of *conserved adjacencies*, i.e. those adjacencies that are common to two genomes, with the restriction that the gene content of both genomes is identical [44, 60]. Thereby, the number of conserved adjacencies constitutes the dual measure of the number of breakpoints between both sequences [47].

With the adoption of gene families, gene duplicates are introduced, i.e., the occurrence of several members of the same family in one genome [45, 49]. Gene duplicates allow for multiple scenarios of ancestral gene order. One possibility to resolve the consequential ambiguities consists in computing a matching between orthologous subsets of given family members, with some predefined constraints on the structure of the matching. This general principle, which relates also to ortholog identification [20], was introduced by David Sankoff with the notion of *exemplar distance* [45], where the main ortholog (the exemplar) of each family is kept. This initial model was later generalized to less constrained classes of matchings where one or more genes per family is kept, always leading to NP-hard computational problems [2, 11, 66], although practically efficient solutions were designed, using heuristics [29] or integer linear programming [1].

Family-free adjacencies. Recently, a gene family-free model was introduced to compute the number of conserved adjacencies in pairwise comparison [22]. The computational problem being NP-hard, exact and heuristic algorithms were presented with feasible running times in practice. In this section, we advance towards a more general model applicable for the simultaneous study of several genomes. Conserved adjacencies obtained in this approach can further benefit ancestral genome reconstruction, as it will be explained in Section 5.

The genome model described in Section 2 is neither restricted to one-to-one relations between genes, nor to closed sets of gene family members. In the subsequent analysis, unconnected genes are omitted from the chromosomal sequences. The remaining genes form connected components of size two or larger. Their size is typically greater than their gene family counterparts. Further, opposing the gene family concept, these connected components are not required to equal their transitive closure.

Given $k \geq 2$ genomes, we aim to find a matching between genes, analogous to previous family-based approaches [1, 10, 45]. One way is to find all completely connected subgraphs of size k in the gene similarity graph and then perform a k-dimensional matching (also known as k-matching). Yet, this approach eliminates many connected components that do not form complete cliques or spread over only a smaller subset of genomes. Consequently, with increasing number of genomes in the dataset, the matching size will decrease until only few fully connected genes remain. In this work we use a *partial k-matching* which allows for missing genes and edges:

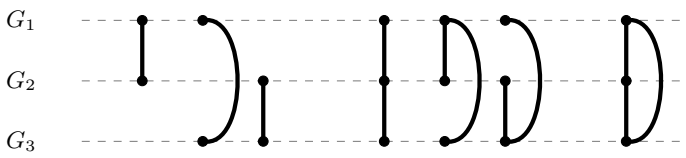


Fig. 2. The 7 valid types of components of a partial 3-matching

Definition 2 (Partial *k*-matching). *Given a gene similarity graph* $B = (G_1, \ldots, G_k, E)$*, a* partial k-matching $\mathcal{M} \subseteq E$ *is a selection of edges such that for each connected component* $C \subseteq B_{\mathcal{M}} := (G_1, \ldots, G_k, \mathcal{M})$ *no two genes in* C *belong to the same genome.*

Figure 2 depicts all valid types of components in a partial k-matching for $k = 3$. The partial k-matching is closely related to the *intermediate matching* [1] for $k = 2$. Just as in the latter, a partial k-matching can saturate an arbitrary number of edges of the initial k-partite graph B but differs in that it is not required to saturate at least one edge per connected component. Our motivation to reject this constraint is discussed further below.

Biological interpretation. Relating to the underlying mechanism of gene family evolution, a connected component in the partial k-matching represents a tentative sub-family assignment in which two intrinsic aspects of gene family prediction are addressed; first, the similarity measure between genes is generally not transitive; second, genes and gene families may arise or vanish along the evolutionary process whereas some genes that are intermittently indispensable for the organism emerge as main orthologs. The biological interpretation of the matching is limited by the restriction to one-to-one assignments between genes and by the fact that the matching does not consider the underlying phylogeny of species and thus is unable to differentiate between orthologs and paralogs. As such, our method is susceptible to non-ortholog assignments in entangled events of gene deletions. Thus it is deceptive to relate a connected component in the partial k-matching to an ortholog assignment. Rather, under the optimization problem stated further below, it represents a tentative sub-family determined by the most parsimonious homology assignment with respect to gene similarity and gene order. Conserved subsets of outparalogs are likely to be assigned to the same connected components, whereas for inparalogs the main orthologs are likely to be matched.

Constructing a partial k-matching. We assume for now that a partial k-matching $\mathcal{M}$ is given. For any two genomes G and H in the gene similarity graph we define $\mathcal{M}_{GH} \subseteq \mathcal{M}$ as the set of matched edges between G and H. We call a gene *GH-saturated* if it is incident to an edge in $\mathcal{M}_{GH}$. Two GH-saturated genes are *consecutive* with respect to G and H if no GH-saturated gene lies between them. Further, two pairs of consecutive GH-saturated genes (g, g') in genome G, with g to the left of g', and (h, h') in genome H, form a *conserved adjacency* if

(a) for h left of h' in H, $sgn(g) = sgn(h)$ and $sgn(g') = sgn(h')$ or
(b) for h right of h' in H, $sgn(g) \neq sgn(h)$ and $sgn(g') \neq sgn(h')$,

where the orientation of a gene (or telomere) g is determined by the following function:

$$sgn(g) = \begin{cases} 1 & \text{if } g \text{ is in forward direction} \\ -1 & \text{if } g \text{ is in backward direction} \\ 0 & \text{if } g \text{ is a telomere} \end{cases}$$

For example, the consecutive pair of genes $(2, \overline{3})$ and $(6, \overline{7})$ in Figure 1(b) represent a conserved adjacency. Following [22], we define a scoring scheme for adjacencies:

$$s(g, g', h, h') = \begin{cases} \sqrt{w(e_{g,h}) \cdot w(e_{g',h'})} & \text{if } (g, g'), (h, h') \text{ form a cons. adjacency} \\ 0 & \text{otherwise} \end{cases}$$

The convex nature of the scoring scheme rewards conserved adjacencies between high weighted edges the most, whereas combinations of high and low weighted, or low weighted edges are decreasingly scored. While a matching that creates many conserved adjacencies is often more appreciated than a matching with few conserved adjacencies, maximizing the number of conserved adjacencies is not desirable at any price. For example, the matching depicted in Figure 1(b) contains an adjacency between genes $(\overline{3}, 4)$ and $(7, 8)$ at the expense of dismissing the stronger edge between genes $(4, 8)$, which is selected in the matching displayed in Figure 1(c). Hence we view a matching as a trade-off between two competing properties, namely similarity and synteny. We quantify both in a matching $\mathcal{M}$ between genomes $\mathcal{G} = \{G_1, \ldots, G_k\}$ by means of the following measures:

$$adj(\mathcal{M}) = \sum_{G,H \in \mathcal{G}} \sum_{\substack{g \text{ left of } g' \text{ in } G \\ h,h' \text{ in } H}} s(g, g', h, h') \tag{1}$$

$$edg(\mathcal{M}) = \sum_{e \in \mathcal{M}} w(e) \tag{2}$$

Extending [22], we propose to find a partial k-matching that maximizes a linear combination of both quantities:

Problem 1 (FF-Adjacencies). Given a gene similarity graph $B = (G_1, \ldots, G_k, E)$ and some $\alpha \in [0, 1]$, find a partial k-matching $\mathcal{M}$ such that the following formula is maximized:

$$\mathcal{F}_\alpha(\mathcal{M}) = \alpha \cdot adj(\mathcal{M}) \; + \; (1 - \alpha) \cdot edg(\mathcal{M}). \tag{3}$$

Thereby α is a user-controlled parameter that can be adjusted in favor of similarity or synteny.

Rejection of intermediate matching constraints. Recall that a partial k-matching for $k = 2$ differs from the intermediate matching only by omitting the constraint that for each connected component at least one edge must be matched. While such restriction is reasonable in gene family studies, where family assignments act as filter in reducing false positive associations between genes, the gene similarity graph can include also small weakly connected components (depending on the particular similarity function) that most likely represent false positives. Substituting the intermediate matching which was used in the initial gene family-free approach [22] for the partial k-matching may have a crucial effect on α in solving Problem FF-Adjacencies. While in pairwise comparison where $\alpha = 0$, both matchings coincide, the choice of edges in the intermediate matching is increasingly limited, when $\alpha > 0$. Discarding the constraint of keeping at least one edge per connected component allows more freedom in the choice of edges included in the matching and thus may lower the number of false positive assignments. However, it does so at the cost of increasing the combinatorial solution space that must be explored in solving Problem FF-Adjacencies. That is because the constraints of the intermediate matching enable the reduction of the solution space by identifying anchors in the gene similarity graph. Using a partial k-matching, we lack sensible constraints of the matching that can be exploited to identify anchors beforehand. Nevertheless, heuristic methods can be applied to establish anchors based on highly conserved structures in the gene similarity graph that are likely preserved in optimal solutions of Problem FF-Adjacencies. These methods will not be discussed here.

3.2 Common Intervals

The concept of *common intervals* is used to represent two or more genomic segments (usually from different genomes) that are composed of the same set of genes. The presence of such segments in the genomes of different species suggests either functional coupling of the involved genes, as observed in operons in prokaryotes, or remnant ancestral gene order, often referred to as *syntenic blocks*, which are used to study large-scale genome evolution. Over the past years, the common intervals model has been generalized to increase its applicability: Starting from a model that requires genomes to be permutations of each other [32, 33, 56], it extended to a sequence-based model that allows multiple occurrences of the same gene and differences in the gene composition of genomes [21, 50]. Finally it was redefined in different ways to account for small differences in the gene content of otherwise well-conserved segments. The most notable of the latter extensions are *r-windows* [23], *max-gap clusters* [4, 31] and *approximate common intervals* [12, 34, 43].

Currently, all approaches to common interval detection require as a prerequisite that the genes of the studied genomes are partitioned into gene families. It is evident that errors in this assignment can have a negative impact on common intervals detection. In the classical common intervals model a single unrecognized homology can prematurely end a conserved segment, or even cause the whole segment to remain unrecognized. Approximate common intervals are to some

extent robust against errors in gene family assignment. An unrecognized homology between two genes may be interpreted as a combined gene insertion/gene deletion. However, in presence of a large number of erroneous gene family assignments this workaround quickly reaches its limits. Another drawback of the current approach is that all information on alignment scores is discarded once gene families are assigned, such that later on, it makes no difference if two genes that are each others' counterpart in a pair of common intervals are strong bidirectional best hits or barely made it into the same gene family and may not even be true homologs after all.

To make better use of positional information and pairwise gene similarity scores, we can use a partial k-matching, as introduced earlier in this section, and simply translate each connected component into one gene family. (Strictly speaking, these are rather sub-families, as discussed previously.) However, conserved adjacencies, the only type of positional information currently used to obtain partial k-matchings, are not optimal in the context of common intervals detection. Typically their definition allows for unrestricted internal rearrangements and disregards gene orientation. The rationale behind this approach is not that conservation of gene order and orientation are supposed to be meaningless, but merely that it is difficult to decide *ad hoc* how much internal rearrangement in a conserved segment is plausible. In practice, a post-processing step can be applied to screen the predicted conserved segments for these qualities. A more integrative approach are *generalized adjacency clusters* which employ a user-defined parameter to restrict internal rearrangements [67].

The above considerations suggest that for common intervals more suitable positional information for gene family assignment could be obtained if the partial k-matching was not only based on conserved adjacencies, but the conserved neighborhood of up to $\theta > 0$ genes to the left and right of each gene. To obtain such a matching, we introduce the notion of θ-*neighbors*: Two genes g and g' in genome G are θ-neighbors with respect to G and H if at most $\theta - 1$ GH-saturated genes lie between them. Two pairs of θ-neighbors (g, g') in genome G and (h, h') in genome H form a θ-*adjacency* if the corresponding edges $e_{g,h}$ and $e_{g',h'}$ are part of $\mathcal{M}_{GH}$. An initial scoring scheme for θ-adjacencies could look as follows:

$$s^{\theta}(g, g', h, h') = \begin{cases} \sqrt{w(e_{g,h}) \cdot w(e_{g',h'})} & \text{if } (g, g') \text{ and } (h, h') \text{ form a } \theta\text{-adjacency} \\ 0 & \text{otherwise} \end{cases}$$

It can be extended by a weighting scheme that values pairs of θ-neighbors the higher the closer they are.

While the use of positional information is most likely an advantage for gene family assignment, the restriction of gene families to at most one gene per genome, a consequence of the partial k-matching, is clearly not. In fact, it is not only unnecessary but even unwanted in common intervals detection. It prevents the detection of duplicate occurrences of genes within a common interval, as well as multiple occurrences of common intervals in a genome. Both findings are certainly interesting as they hint at segmental or whole genome duplications.

In the remainder of this section, we broach a gene family-free approach for common intervals detection that avoids the above mentioned restrictions. We first study the case of two genomes G and H. Any pair of intervals (I, J) on G and H can be common intervals. Therefore we build for each (I, J) a maximum weighted bipartite matching $\mathcal{M}_{I,J}$ between the gene sets of I and J. This is equivalent to solving Problem FF-Adjacencies with $\alpha = 0$ for $G_1 = I$ and $G_2 = J$.

An unmatched gene in I and J is either a duplicate occurrence if it is incident to an unchosen edge within the interval pair, or an inserted gene, if there are no incident edges or all of them point to a gene outside the interval pair. We obtain a matching score $score(\mathcal{M}_{I,J}) = \mathcal{F}_0(\mathcal{M}_{I,J})$ that needs to be corrected for the number of genes occurring in the intervals. Otherwise, the biggest score is obtained for (G, H), the interval pair defined by the complete genomes. Simply normalizing $score(\mathcal{M}_{I,J})$ by the length of I and J is also not advisable, as it causes the best-scoring common intervals to be of length one, the best scoring pair of genes. Instead a trade-off between matching score and interval compactness needs to be defined. The corrected score can then be used to decide whether an interval pair should pass for a conserved segment or not. For $k > 2$ genomes, the matching score can be defined as the sum over all pairwise matching scores which equals the score of a partial k-matching over all genomes.

The computation of a single matching $\mathcal{M}_{I,J}$ can be done in $O(\max\{|I|, |J|\}^3)$ time using the Hungarian Method [35]. However, already for two genomes there are $O(|G|^2|H|^2)$ interval combinations that need to be tested. One order of magnitude is saved if the initial definition of common intervals is used that neither allows duplicate genes nor gene insertions/deletions. In this case, only intervals of the same size need to be paired. For larger k, the complexity increases further, as all $O(k^2)$ pairwise genome combinations need to be considered. With polynomials of such high degrees in the asymptotic time complexity, it remains to be seen to what extent matching-based approaches are feasible in practice.

4 Genome Rearrangements

The study of genome rearrangements leads to a better understanding of the dynamics of genome structure over time. Typical rearrangement operations are the *inversion* of a piece of a chromosome, the *translocation* of material between two chromosomes, or the *fusion* and *fission* of chromosomes. These operations are explicitly modeling the modification of the genome over time and the methods therefore are called *rearrangement model-based* [30, 44, 63], in contrast to the *rearrangement model-free* methods that we discussed in the previous section, which only study and compare static properties of the genomes.

In rearrangement model-based methods, given two genomes and a set of rearrangement operations, two problem variants are typically considered: (1) calculate the minimum number of steps that are necessary to transform one genome into another, the so-called *genomic distance problem*, and (2) find a series of operations that perform such a transformation, the *genomic sorting*

problem. Traditional approaches to analyze these problems are family-based, and the vast majority of methods also adopt the simplifying assumption that exactly one occurrence of each family appears in each genome, which allows the existence of several polynomially-time computable methods, including for the popular Double Cut and Join (DCJ) rearrangement model [7, 63].

While the sorting problem, especially for the case of multiple genomes and their relation along the branches of a phylogenetic tree, will be addressed briefly in the following Section 5, here we concentrate on distance calculations in a family-free setting. In general, similarly to the rearrangement model-free measure of conserved adjacencies described in Section 3, the challenge is finding pairwise gene assignments based on similarities while minimizing the distance. In the following we will sketch a natural modification of existing approaches for the DCJ model. Whether this will lead to meaningful distances and allows for efficient algorithms has yet to be shown.

4.1 The Weighted Adjacency Graph

Recall that a gene is an oriented interval of a chromosome. We now represent a gene by the two extremities of its interval, called *tail* and *head*. The tail of gene g is denoted by g^t and the head by g^h. In a family-based setting composed of n gene families, consider that each one of two genomes G and H has exactly n genes, one occurrence of each family. A data structure that has proven to be useful in the study of the DCJ rearrangement model in this context is the adjacency graph $AG(G, H)$. This graph has a vertex for each adjacency of either of the two given genomes, and for each one of the two extremities of each gene there is an edge connecting the two vertices, one in G and the other in H, that contain this extremity. The graph is bipartite and a collection of paths and cycles, because each vertex has either degree one or degree two. The DCJ rearrangement distance can easily be calculated from this graph using the formula $d_{DCJ} = n - c - i/2$, where c is the number of cycles and i is the number of paths with an odd number of edges in $AG(G, H)$ [7]. Since, in the linear unichromosomal case that we consider in this paper, the adjacency graph has exactly two paths and otherwise only cycles, $i/2$ is either 0 or 1. Therefore, the similarity of two genomes G and H is closely related to the number of cycles in the adjacency graph $AG(G, H)$.

While the original adjacency graph clearly depends on the assignment of gene families, we observe that based on the information in the gene similarity graph from Section 2 we can obtain a data structure that resembles some of the properties of the adjacency graph. This new data structure might thus be a good basis for DCJ-like rearrangement distance calculations in a family-free setting:

Definition 3 (Weighted Adjacency Graph). *The* weighted adjacency graph *$WAG(G, H)$ of two genomes G and H has a vertex for each adjacency in G and a vertex for each adjacency in H. For a gene g in G and a gene h in H with similarity $\sigma(g, h) > 0$ there is one edge connecting the vertices containing the two heads g^h and h^h and one edge connecting the vertices containing the two tails g^t and h^t. The weight of each of these edges is $w(e_{g,h}) := \sigma(g, h)$.*

As an example, the gene similarity graph for the two genomes $G = (\circ\ \overline{1}\ 2\ \overline{3}\ 4\ \circ)$ and $H = (\circ\ \overline{5}\ 6\ 7\ 8\ 9\ \circ)$ and six edges with non-zero weight, and the corresponding weighted adjacency graph are given in Figure 3.

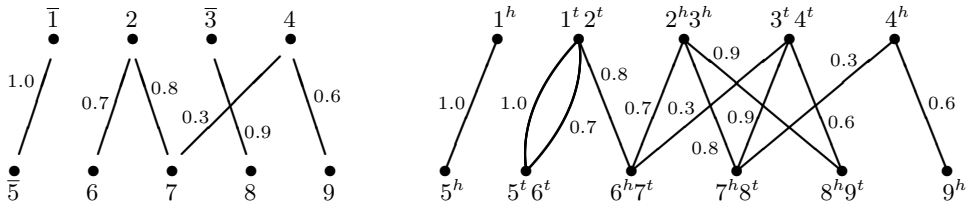


Fig. 3. Gene similarity graph (left) and the resulting weighted adjacency graph $WAG(G, H)$ (right) for two genomes $G = (\circ\ \overline{1}\ 2\ \overline{3}\ 4\ \circ)$ and $H = (\circ\ \overline{5}\ 6\ 7\ 8\ 9\ \circ)$

Note that if G and H have the same number of genes and the similarity measure σ forms a perfect matching with weight 1 for all edges of the matching and weight 0 otherwise, then the weighted adjacency graph reduces to the ordinary adjacency graph.

4.2 The Weighted Double-Cut-and-Join Distance

As for the case of conserved adjacencies, where instead of the breakpoint distance we calculate a matching maximizing an adjacency score in Equation (3), here we first define a similarity measure that, if needed, can easily be converted into a distance.

Again, the similarity measure is based on a matching $\mathcal{M}$ of the genes in G and the genes in H. Let $\mathcal{I}(G, H; \mathcal{M})$ be a graph derived from the weighted adjacency graph $WAG(G, H)$ and the matching $\mathcal{M}$ by first removing from $WAG(G, H)$ each unmatched gene, consequently merging the two vertices containing its extremities, and second keeping only the edges representing extremities of gene pairs from $\mathcal{M}$. This graph has the shape of a standard adjacency graph and thus is a collection of cycles and paths. We denote by $\mathcal{C}(\mathcal{M}) \equiv \mathcal{C}(G, H; \mathcal{M})$ the set of connected components of $\mathcal{I}(G, H; \mathcal{M})$.

The graph derived from the weighted adjacency graph of Figure 3 and the matching $\mathcal{M} = \{(1, 5), (2, 6), (3, 8), (4, 9)\}$ is given in Figure 4.

Since we know that the number of DCJ operations is closely related to the number of cycles in the adjacency graph, we define a score function whose domain is defined by gene similarities and cycles in the matching. Therefore, in analogy to the corresponding formula for conserved adjacencies in Equation (3), we propose the following objective function:

$$\mathcal{F}_{\alpha}^{DCJ}(\mathcal{M}) = \alpha \cdot cyc(\mathcal{M}) + (1 - \alpha) \cdot edg(\mathcal{M})$$

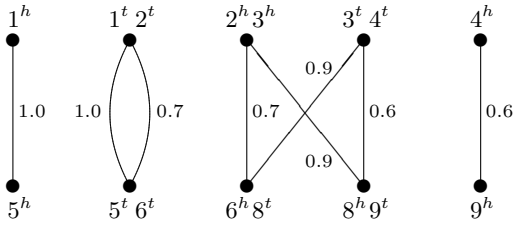


Fig. 4. The graph derived from the weighted adjacency graph of Figure 3 and the matching $\mathcal{M} = \{(1,5),(2,6),(3,8),(4,9)\}$

where

$$cyc(\mathcal{M}) = \sum_{C \in \mathcal{C}(\mathcal{M})} \left(\frac{1}{|C|} \sum_{e \in C} w(e) \right)$$

and $edg(\mathcal{M})$ is the same as in Equation (2). Again, $\alpha \in [0,1]$ is a parameter that allows to balance between the two extremes, here between rearrangements ($\alpha = 1$) and gene similarities ($\alpha = 0$). Nevertheless, even for $\alpha = 1$ gene similarities are not ignored since the weights $w(e)$ also form an essential part of the cycle score $cyc(\mathcal{M})$. Note that the normalization $1/|C|$ in $cyc(\mathcal{M})$ is designed such that many short cycles are preferred over fewer long ones. For example, if all edges have the same weight w, two cycles of length 2 receive the score $2w$, which is twice the score of one cycle of length 4. The cycle score of the graph shown in Figure 4 is $cyc(\mathcal{M}) = \frac{1}{1}\cdot 1.0 + \frac{1}{2}\cdot(1.0+0.7) + \frac{1}{4}\cdot(0.7+0.9+0.6+0.9) + \frac{1}{1}\cdot 0.6 = 3.225$.

It is unlikely to find an efficient algorithm to compute a matching $\mathcal{M}$ that maximizes $\mathcal{F}_\alpha^{DCJ}(\mathcal{M})$, but the solution of this optimization problem through integer linear programming seems possible and will be the subject of further research.

It is also an open question how to treat genes that are not covered by $\mathcal{M}$. They can be explained as being inserted or deleted during the course of evolution. Thus, a more general score function might consider these genes and prefer sorting scenarios with a low number of insertion/deletion events, similar to existing family-based approaches [14, 25].

Even further reaching might be approaches that do not rely on any matching, and instead optimize an objective directly defined on the weighted adjacency graph, for example a weighted version of maximum cycle decomposition.

5 Ancestral Genome Reconstruction

Studying conservation of gene order or rearrangement processes in the light of a phylogeny – given or unknown – can provide deeper insight into evolutionary mechanisms, gene functions, or the phylogeny itself. In this section, we will discuss how a partial k-matching can be used for ancestral genome reconstruction.

Phylogeny aware optimization. A natural first step when reconstructing ancestral gene orders is to take phylogenetic information into account. Apart from ancestral reconstruction, this can actually be done in general to improve the construction of the partial k-matching. Given an edge-weighted phylogenetic tree, say $\mathcal{T}$, for the species under consideration where the edge weights reflect the phylogenetic/evolutionary distance, the lengths of the paths between all pairs of species define an additive distance matrix $D^{\mathcal{T}}$. As additivity gives a one-to-one correspondence of $D^{\mathcal{T}}$ and $\mathcal{T}$, including the pairwise distances into the optimization implicitly also includes the topology of $\mathcal{T}$. These distances can be used to scale the pairwise scores in the objective function – close relatives receive a higher score than more distant pairs:

$$\begin{aligned}\mathcal{F}_{\alpha,\mathcal{T}}(\mathcal{M}) &= \alpha \cdot \sum_{G,H}(D^{\mathcal{T}}_{max} - D^{\mathcal{T}}_{GH})\, adj(\mathcal{M}_{GH}) \\ &\quad + (1-\alpha) \cdot \sum_{G,H}(D^{\mathcal{T}}_{max} - D^{\mathcal{T}}_{GH})\, edg(\mathcal{M}_{GH}) \\ &= \sum_{G,H}(D^{\mathcal{T}}_{max} - D^{\mathcal{T}}_{GH})\,(\alpha \cdot adj(\mathcal{M}_{GH}) + (1-\alpha) \cdot edg(\mathcal{M}_{GH}))\end{aligned}$$

where

$$D^{\mathcal{T}}_{max} = \max_{G,H}\{D^{\mathcal{T}}_{GH}\}.$$

Ancestral genes. To be able to reconstruct ancestral gene orders, we first need to define ancestral genes and the ancestral gene content of ancestral genomes. To this end, we leave the family-free approach and rely on the assignments given by the partial k-matching. From such assignments, gene families can be derived by simply assigning all genes from a connected component in a partial k-matching to one family. As mentioned in Section 3.2, strictly speaking, these are rather gene sub-families. Recall further that the partial k-matching is defined such that within each connected component formed by saturated edges no two genes belong to the same genome. If all components are k-cliques, then genomes can be modeled as signed permutations. But in general, components might cover less than k genomes, i.e., not all genomes have the same gene content, although genomes do not have duplicated genes, thus leading to easier problems.

Based on the gene sub-families, we can infer the ancestral gene content from standard methods [18] or methods tailored for genome rearrangement problems [24, 53].

Ancestral gene orders. Similarly to the computation of genomic distances (Section 4), the reconstruction of ancestral gene orders can be seen from two points of view – incorporating a rearrangement model-based approach or not. Once gene families have been defined from the partial k-matching, we have the gene orders of the extant genomes. Thus, we can apply rearrangement model-based methods allowing for unequal gene content such as [24, 49, 53]. Usually, such methods,

following a parsimony approach, would aim at minimizing the total number of operations along the tree edges, which in most cases will lead to computationally hard optimization problems.

In the rearrangement model-free approach, ancestral syntenic characters are determined which induce a (partial) gene order. In our case, adjacencies qualify as ancestral syntenic characters. The remaining questions are then (1) how to infer the ancestral adjacencies, and (2) whether a set of adjacencies assigned to an ancestral node is concordant with some valid gene order, i.e., a collection of linear (and circular) chromosomes where each gene has at most two neighbors.

For a median-of-three, the above questions can easily be answered. Following a parsimony approach, the 0/1-assignment of an adjacency to the median boils down to a majority vote. Further, in almost all rearrangement median models, any adjacency present in at least two genomes is contained in any optimal median. In the case of signed gene orders, this selection will always ensure compatibility with a collection of linear and circular gene orders, and the inferred partial k-matching defines implicitly a set of linear or circular genome segments. Note however that this median genome might not be optimal for a given rearrangement model; however, it is a valid set of ancestral genome segments that has been inferred in a joint process, together with putative gene sub-families.

For general trees, one could follow rearrangement model-free approaches that try to find a most parsimonious labeling of the whole tree that is at the same time consistent with some linear or circular gene order [52], or one could concentrate on a single ancestral node as, e.g., done in several recent works [17,37]. The method by Chauve and Tannier [17] relies on the Dollo principle, where only adjacencies conserved in pairs of genomes whose path in the species tree contain that ancestor are deemed ancestral; other approaches can select or score adjacencies using a Fitch principle [37].

The Dollo principle can easily be included into the optimization of the partial k-matching by introducing a factor π^A_{GH} that equals one if the path between G and H contains the ancestor A and zero otherwise:

$$\mathcal{F}_{\alpha,\mathcal{T},A}(\mathcal{M}) = \sum_{G,H} \pi^A_{GH} \left(D^{\mathcal{T}}_{max} - D^{\mathcal{T}}_{GH}\right) \left(\alpha\, adj(\mathcal{M}_{GH}) + (1-\alpha)\, edg(\mathcal{M}_{GH})\right).$$

Thus, adding this feature to the objective function allows to select a set of putative ancestral adjacencies that can also receive a phylogenetic score as we described it earlier. Then existing methods that select a subset of adjacencies that form a valid genome can be used (see [38] for an example).

In this section we outlined how the family-free principle can fit quite naturally in existing approaches to reconstruct ancestral gene orders. This preliminary study opens several interesting research avenues. For example, it is worth to mention that rearrangement model-free reconstruction methods can utilize larger conserved structures than just adjacencies. Thus, e.g., common intervals could be

included by integrating the scoring for θ-adjacencies as proposed in Section 3.2. Also, progressing toward a fully integrated inference process, it would be natural to incorporate the constraints posed by the structure of an ancestral genome; with adjacencies, this reduces to ensuring that every ancestral gene has at most two adjacent neighboring genes. However, integrating such constraints – even if only for a single internal node of a species tree (ancestral genome) – seems to be very challenging. Finally, it would also be interesting to move on from reconstructing the states of the internal nodes of a given phylogeny (the small phylogeny problem) to reconstructing the tree itself. It is known that using gene order data for phylogenetic reconstructions can be more accurate and robust than sequence based methods since they are not affected by gene-tree species-tree issues and less affected by small sequence or alignment errors. Not relying on purely sequence based homology assignments could be a benefit for such reconstructions.

6 Discussion

In this paper we have outlined the potential of family-free methods in various aspects of genome comparison. Gene families are generally computationally predicted and serve as basis for a large variety of current comparative genomics studies. Since the predicted families may not be concordant with the underlying true biological gene families, erroneous gene family assignments can deteriorate subsequent analyses. Most importantly, comparative genomics methods require prior gene family assignments, yet the attained information about the structural organization of the genome may in turn actually help to improve the initially required gene family assignments. Consequently we propose the use of a *gene similarity graph* as underlying data structure in genome comparison. Therein genes are associated with each other by weighted edges according to a normalized similarity measure. In practice, sequence similarity scores can be employed in constructing the graph.

The underlying strategy of almost all presented methods is tantalizingly simple and boils down to obtain a one-to-one matching between orthologous genes of the gene similarity graph by solving an optimization problem. More specifically, a linear combination of a synteny (or rearrangement) score and a similarity score, parameterized by α, between saturated genes is optimized. Here, we give users the choice in favoring one of the two quantities over the other by adjusting α in each particular analysis. At this point, we like to acknowledge an inherent disadvantage of a one-to-one matching, namely its inability to account for inparalogous genes. Thus, the detection of inparalogs remains part of post-processing steps which identify unsaturated genes with high similarities to other genes of the same genome.

In Section 3 we studied two forms of conserved structures: adjacencies and common intervals. In the former, we generalized the problem of family-free computation of adjacencies of [22], called FF-Adjacencies, towards the simultaneous study of more than two genomes. Thereby we introduced the notion of a *partial*

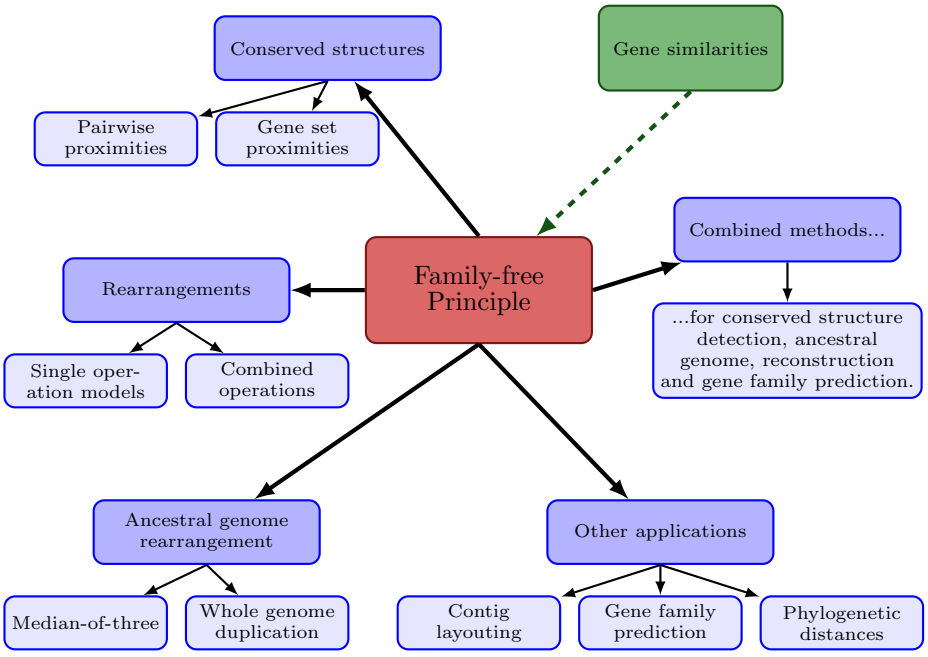


Fig. 5. Various fields of comparative genomics can be explored under the family-free model such as conserved structure detection or reconstruction of ancestral genomes, employing different gene similarity measures (based on alignment scores, functional similarity, etc).

k-matching, which allows to incorporate in solutions of Problem FF-Adjacencies sparsely interconnected genes as well as connected components that are only contained in subsets of the genomes. We also discussed two possible approaches towards family-free common intervals by introducing a scoring scheme for *θ-adjacencies*, which is a co-localization measure for genes similar to adjacencies. We further outlined a more dynamic, but also computationally more expensive approach based on performing local maximum matchings. Complementing the study of conserved structures, we turned in Section 4 to model-based genome comparison by introducing the *weighted adjacency graph*. On this basis we proposed a *weighted DCJ* distance following a similar strategy as in the previous section. We further showed in Section 5 how the reconstruction of ancestral genomes can be performed using the family-free principle. Thereby we studied the concept of family-free adjacencies in a phylogeny-aware setting using existing approaches of reconstructing ancestral gene orders.

This work presents a number of initial studies in a new field of genome comparison which aims at developing methods where prior gene family assignments are no longer required. It consequently offers many directions in which these

studies can be extended (see Figure 5). Most evidently, the principle of family-free genome comparison can be applied to the numerous existing family-based studies. More interestingly, the family-free principle could even be integrated into a methodology for joint inference of gene families, conserved structures and ancestral gene orders at the same time, extending presented work in reconstructing ancestral gene orders. Even though such venture most likely involves a more complex data structure and a potentially increased solution space, the question remains unanswered if the stronger signal gained from harvesting more information from the genomic datasets may reduce the computational cost in finding optimal solutions. Finally, it is worth to mention that the family-free principle may be particularly beneficial in studying partially sequenced (or assembled) genomes, as methods in gene family prediction tend to be susceptible for missing genes. Here, the family-free approach can offer improvements for inferring phylogenetic distances of incomplete genomes, but also in detecting conserved structures, which may lead to improved methods in contig layouting.

While sequence similarity between genes is an obvious and reasonable measure in constructing the gene similarity graph, similarity scores can also integrate additional information such as functional similarity. Such information can be obtained from various databases, most notably, from the Gene Ontology database [3]. Family-free genome comparisons of this kind may give further insights into the functional organization of the genome.

Acknowledgements

MDVB is funded by the Brazilian research agency CNPq grant PROMETRO 563087/10-2. DD receives a scholarship from the CLIB Graduate Cluster Industrial Biotechnology. KJ is funded by DFG grant ST 431/5-1. AT is a research fellow of the Alexander von Humboldt Foundation.

References

1. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: Efficient tools for computing the number of breakpoints and the number of adjacencies between two genomes with duplicate genes. J. Comp. Biol. 15(8), 1093–1115 (2008)
2. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: On the approximability of comparing genomes with duplicates. J. Graph Algorithms Appl. 13(1), 19–53 (2009)
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat. Genet. 25(1), 25–29 (2000)
4. Bergeron, A., Corteel, S., Raffinot, M.: The algorithmic of gene teams. In: Proceedings of WABI 2002. LNCS, vol. 2452, pp. 464–476 (2002)
5. Bergeron, A., Mixtacki, J., Stoye, J.: On sorting by translocations. J. Comp. Biol. 13(2), 567–578 (2006)

6. Bergeron, A., Stoye, J.: On the similarity of sets of permutations and its applications to genome comparison. J. Comp. Biol. 13(7), 1340–1354 (2006)
7. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Proceedings of WABI 2006. LNBI, vol. 4175, pp. 163–173 (2006)
8. Bernt, M., Merkle, D., Middendorf, M.: Solving the preserving reversal median problem. IEEE/ACM Trans. Comput. Biol. Bioinf. 5, 332–347 (2008)
9. Blin, G., Chateau, A., Chauve, C., Gingras, Y.: Inferring positional homologs with common intervals of sequences. In: Proceedings of RECOMB-CG 2006. pp. 24–38. Springer (2006)
10. Blin, G., Chauve, C., Fertin, G.: The breakpoint distance for signed sequences. In: Proceedings of CompBioNets 2004. Texts in Algorithmics, vol. 3, pp. 3–16 (2004)
11. Blin, G., Chauve, C., Fertin, G., Rizzi, R., Vialette, S.: Comparing genomes with duplications: A computational complexity point of view. IEEE/ACM Trans. Comput. Biology Bioinf. 4(4), 523–534 (2007)
12. Böcker, S., Jahn, K., Mixtacki, J., Stoye, J.: Computation of median gene clusters. J. Comput. Biol. 16(8), 1085–1099 (2009)
13. Bourque, G., Pevzner, P.A.: Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Res. 12(1), 26–36 (2002)
14. Braga, M.D.V., Willing, E., Stoye, J.: Double cut and join with insertions and deletions. J. Comp. Biol. 18(9), 1167–1184 (2011)
15. Caprara, A.: The reversal median problem. INFORMS J. Computing 15(1), 93–113 (2003)
16. Chauve, C., El-Mabrouk, N., Gueguen, L., Semeria, M., Tannier, E.: Duplication, Rearrangement and Reconciliation: a follow-up 13 years later (2013), in this volume
17. Chauve, C., Tannier, E.: A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Comp. Biol. 4(11), e1000234 (2008)
18. Csurös, M.: Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26(15), 1910–1912 (2010)
19. Darling, A.E., Mau, B., Perna, N.T.: progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. PloS one 5(6), e11147 (2010)
20. Dewey, C.N.: Positional orthology: putting genomic evolutionary relationships into context. Brief. Bioinform. 12(5), 401–412 (2011)
21. Didier, G., Schmidt, T., Stoye, J., Tsur, D.: Character sets of strings. J. Discr. Alg. 5(2), 330–340 (2007)
22. Doerr, D., Thévenin, A., Stoye, J.: Gene family assignment-free comparative genomics. BMC Bioinformatics 13(Suppl 19), S3 (2012)
23. Durand, D., Sankoff, D.: Tests for gene clustering. J. Comp. Biol. 10, 453–482 (2003)
24. Earnest-DeYoung, J.V., Lerat, E., Moret, B.M.E.: Reversing gene erosion - reconstructing ancestral bacterial genomes from gene-content and order data. In: Proceedings of WABI 2004. LNCS, vol. 3240, pp. 1–13 (2004)
25. El-Mabrouk, N.: Sorting signed permutations by reversals and insertions/deletions of contiguous segments. J. Discrete Alg. 1(1), 105–122 (2001)
26. Feijão, P., Meidanis, J.: SCJ: A breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Trans. Comput. Biol. Bioinf. 8(5), 1318–1329 (2011)
27. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: Combinatorics of Genome Rearrangements. MIT Press (2009)
28. Frech, C., Chen, N.: Genome-wide comparative gene family classification. PLoS one 5(10), e13409 (2010)

29. Fu, Z., Chen, X., Vacic, V., Nan, P., Zhong, Y., Jiang, T.: MSOAR: A high-throughput ortholog assignment system based on genome rearrangement. J. Comp. Biol. 14(9), 1160–1175 (2007)
30. Hannenhalli, S., Pevzner, P.A.: Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. J. ACM 46(1), 1–27 (1999)
31. He, X., Goldwasser, M.H.: Identifying conserved gene clusters in the presence of homology families. J. Comp. Biol. 12(6), 638–656 (2005)
32. Heber, S., Stoye, J.: Algorithms for finding gene clusters. In: Proceedings of WABI 2001. LNCS, vol. 2149, pp. 252–263 (2001)
33. Heber, S., Mayr, R., Stoye, J.: Common intervals of multiple permutations. Algorithmica 60(2), 175–206 (2011)
34. Jahn, K.: Efficient computation of approximate gene clusters based on reference occurrences. J. Comput. Biol. 18(9), 1255–1274 (2011)
35. Kuhn, H.W.: The hungarian method for the assignment problem. Nav. Res. Logist. Q. 2(1-2), 83–97 (2006)
36. Li, L., Stoeckert, C.J., Roos, D.S.: OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9), 2178–2189 (2003)
37. Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Zhang, L., Miller, W., Haussler, D.: DUPCAR: reconstructing contiguous ancestral regions with duplications. J. Comp. Biol. 15(8), 1007–1027 (2008)
38. Manuch, J., Patterson, M., Wittler, R., Chauve, C., Tannier, E.: Linearization of ancestral multichromosomal genomes. BMC Bioinformatics 13(Suppl 19), S11 (2012)
39. Milinkovitch, M.C., Helaers, R., Depiereux, E., C, T.A., Gabaldon, T.: 2 genomes - depth does matter. Genome Biology 11, R6 (2010)
40. Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L.L.: InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38(Database issue), D196–D203 (2010)
41. Pe'er, I., Shamir, R.: The median problems for breakpoints are NP-complete. Elec. Colloq. on Comput. Complexity 71, 5 (1998)
42. Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T., Jensen, L.J., von Mering, C., Bork, P.: eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. 40(Database issue), D284–D289 (2012)
43. Rahmann, S., Klau, G.W.: Integer linear programs for discovering approximate gene clusters. In: Proceedings of WABI 2006. LNBI, vol. 4175, pp. 298–309 (2006)
44. Sankoff, D.: Edit distances for genome comparisons based on non-local operations. In: Proceedings of CPM 1992. LNCS, vol. 644, pp. 121–135 (1992)
45. Sankoff, D.: Genome rearrangement with gene families. Bioinformatics 15(11), 909–917 (1999)
46. Sankoff, D., Blanchette, M.: Multiple genome rearrangement and breakpoint phylogeny. J. Comp. Biol. 5, 555–570 (1998)
47. Sankoff, D., Blanchette, M.: The median problem for breakpoints in comparative genomics. In: Proceedings of COCOON 1997. LNCS, vol. 1276, pp. 251–263 (1997)
48. Sankoff, D., Cedergren, R., Abel, Y.: Genomic divergence through gene rearrangement. In: Doolittle, R.F. (ed.) Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Meth. Enzymol., vol. 183, chap. 26, pp. 428–438. Academic Press (1990)

49. Sankoff, D., El-Mabrouk, N.: Duplication, rearrangement and reconciliation. In: Sankoff, D., Nadeau, J.H. (eds.) Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families, Computational Biology Series, vol. 1, pp. 537–550. Kluwer Academic Publishers (2000)
50. Schmidt, T., Stoye, J.: Quadratic time algorithms for finding common intervals in two and more sequences. In: Proceedings of CPM 2004. LNCS, vol. 3109, pp. 347–358 (2004)
51. Shi, G., Peng, M.C., Jiang, T.: MultiMSOAR 2.0: An accurate tool to identify ortholog groups among multiple genomes. PLoS one 6(6), e20892 (2011)
52. Stoye, J., Wittler, R.: A unified approach for reconstructing ancient gene clusters. IEEE/ACM Trans. Comput. Biol. Bioinf. 6(3), 387–400 (2009)
53. Tang, J., Moret, B.M., Cui, L., Depamphilis, C.W.: Phylogenetic reconstruction from arbitrary gene-order data. In: Proceedings of BIBE 2004. pp. 592–599. IEEE (2004)
54. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. BMC Bioinformatics 10, 120 (2009)
55. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A.: The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41 (2003)
56. Uno, T., Yagiura, M.: Fast algorithms to enumerate all common intervals of two permutations. Algorithmica 26(2), 290–309 (2000)
57. Wapinski, I., Pfeffer, A., Friedman, N., Regev, A.: Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics 23(13), i549–58 (2007)
58. Wapinski, I., Pfeffer, A., Friedman, N., Regev, A.: Natural history and evolutionary principles of gene duplication in fungi. Nature 449(7158), 54–61 (2007)
59. Waterhouse, R.M., Zdobnov, E.M., Tegenfeldt, F., Li, J., Kriventseva, E.V.: OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Res. 39(Database issue), D283–8 (2011)
60. Watterson, G., Ewens, W.J., Hall, T., Morgan, A.: The chromosome inversion problem. J. Theor. Biol. 99(1), 1–7 (1982)
61. Xu, A.W., Moret, B.M.E.: GASTS: Parsimony scoring under rearrangements. In: Proceedings of WABI 2011. LNBI, vol. 6833, pp. 351–363 (2011)
62. Xu, X., Sankoff, D.: Tests for gene clusters satisfying the generalized adjacency criterion. In: Proceedings of BSB 2008. LNBI, vol. 5167, pp. 152–160 (2008)
63. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21(16), 3340–3346 (2005)
64. Yang, Z., Sankoff, D.: Natural parameter values for generalized gene adjacency. In: Proceedings of RECOMB-CG 2009. LNBI, vol. 5817, pp. 13–23 (2009)
65. Zhang, M., Leong, H.W.: Identifying positional homologs as bidirectional best hits of sequence and gene context similarity. In: Proceedings of ISB 2011. pp. 117–122. IEEE (2011)
66. Zhu, B.: Approximability and fixed-parameter tractability for the exemplar genomic distance problems. In: Proc. of Theory and Applications of Models of Computation. LNCS, vol. 5532, pp. 71–80 (2009)
67. Zhu, Q., Adam, Z., Choi, V., Sankoff, D.: Generalized gene adjacencies, graph bandwidth, and clusters in yeast evolution. IEEE/ACM Trans. Comput. Biol. Bioinf. 6(2), 213–220 (2009)