Chapter 4

Whole-Genome Alignment

Colin N. Dewey

Abstract

Whole-genome alignment (WGA) is the prediction of evolutionary relationships at the nucleotide level between two or more genomes. It combines aspects of both colinear sequence alignment and gene orthology prediction and is typically more challenging to address than either of these tasks due to the size and complexity of whole genomes. Despite the difficulty of this problem, numerous methods have been developed for its solution because WGAs are valuable for genome-wide analyses such as phylogenetic inference, genome annotation, and function prediction. In this chapter, we discuss the meaning and significance of WGA and present an overview of the methods that address it. We also examine the problem of evaluating whole-genome aligners and offer a set of methodological challenges that need to be tackled in order to make most effective use of our rapidly growing databases of whole genomes.

Key words Sequence alignment, Whole-genome alignment, Homology map, Toporthology, Genome evolution, Comparative genomics

1 Introduction

When the problem of biological sequence alignment was first described and addressed in the 1970s, sequencing technology was limited to obtaining the sequences of individual proteins or mRNAs or short genomic intervals. As such, classical sequence alignment (as described in Chapter 7 [1]) is typically focused on predicting homologous positions within two or more relatively short and colinear sequences, allowing for the edit events of substitution, insertion, and deletion. Although limited in its scope, this type of alignment remains extremely important today, with gene-sized alignments forming the basis of most evolutionary studies.

Starting in 1995 with the sequencing of the 1.8 Mb-sized genome of the bacterium *H. influenzae* [2], biologists have had access to a different scale of biological sequences, those of whole genomes. DNA sequencing technology has rapidly improved since that time, and as a result, we have seen an explosion in the availability of whole-genome sequences. As of the writing of this chapter, there are 9071 published complete genome sequences (8380

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_4,

bacterial, 281 archaeal, and 410 eukaryotic), according to the GOLD database [3]. Whole-genome sequencing remains popular, with over 140,000 sequencing projects that are either ongoing or completed.

Along with the ascertainment of these sequences, the problem of whole-genome alignment (WGA) has arisen. As each genome is sequenced, there is interest in aligning it against other available genomes in order to better understand its evolutionary history and, ultimately, the biology of its species. Like classical sequence alignment, WGA is about predicting evolutionarily related sequence positions. However, aligning whole genomes is made more complicated by the fact that genomes undergo large-scale structural changes, such as duplications and rearrangements. In addition, a set of genomes may contain pairs of sequence positions whose evolutionary relationships can be described by any of the three major subclasses of homology: orthology, paralogy, and xenology. As orthologous positions are typically of primary interest, WGA also involves the classification of homologous relationships.

In this chapter, we describe the problem of WGA and the methods that address it. We begin with a thorough definition of the problem and discuss the important downstream applications of WGAs. We then categorize the WGA methods that have been developed and describe the key computational techniques that are used within each category. In addition to describing whole-genome aligners, we also discuss the various approaches that have been used for evaluating the alignments they produce. Lastly, we lay out a number of current methodological challenges for WGA.

2 The Definition and Significance of WGA

2.1 WGA as a Correspondence Between Genomes

In imprecise terms, a WGA is a "correspondence" between genomes. For each segment of a given genome, a WGA tells us where its "corresponding" segments are in other genomes. A segment may be one or more contiguous nucleotide positions within a genome. What does it mean for two genomic segments to "correspond" to each other? In most situations, we consider two segments to be "corresponding" if they are orthologous. Orthologous sequences are those that are evolutionarily related (homologous) and that diverged from their most recent common ancestor (MRCA) due to a speciation event [4]. In contrast, paralogous sequences are homologs that diverged from the MRCA due to a duplication event. Thus, by definition, orthologous sequences are the most closely related pieces of two genomes and, as is more thoroughly discussed later and in Chapter 9 [5], are of primary interest because they are useful for applications such as function prediction and species tree inference. As such, WGA is most commonly taken to be the prediction of orthology between the

components of entire genome sequences. When a WGA also predicts paralogy, typically only paralogs whose MRCA is at least as recent as the MRCA of entire set of genomes are considered, as there is extensive ancient homology within extant genomes.

It is important to note that the orthologous relationships between two genomes do not create a one-to-one correspondence. Duplication events that have occurred since the time of the MRCA of the species can result in a genomic segment in one species having multiple orthologous segments in another. This is a particularly important issue when the genome of one lineage has undergone a whole-genome duplication event since the time of the MRCA. In this situation, few segments of the genome of the nonduplicated lineage have a single ortholog in the other genome.

2.2 Toporthology

In many cases, WGAs do not aim to predict all orthologous sequences. Instead, they only predict toporthology (positional orthology), a distinguished subset of orthology [6, 7]. The concept of toporthology captures the notion that not all orthologous relationships are equivalent in terms of the evolutionary history of the genomic context of the orthologs. Figure 1 gives an example scenario in which toporthology helps to distinguish between two orthologous relationships.

The definition of toporthology relies on a classification of duplication events. A duplication event is considered to be "symmetric" if the removal of either copy of the duplicated genomic material (immediately after the event) reverts the genome to its original (preduplication) state. Examples of symmetric duplications are tandem and whole-genome duplications. If only one specific copy can be removed to undo a duplication event, then the event is considered "asymmetric." In the asymmetric case, the removable copy is referred to as the "target," with the other copy referred to as the "source." Retrotransposition and segmental duplication both belong to the asymmetric class.

With this classification of duplication events in hand, we can now define toporthology. Two genomic segments are toporthologous if they are orthologous and neither segment is derived from the target of an asymmetric duplication event since the time of the MRCA of the segments. Thus, two orthologous segments are toporthologous if their evolutionary history (since the MRCA) only involves symmetric duplication events or asymmetric duplications in which their ancestral segment was part of the source copy.

The important property of toporthologs is that, in the absence of rearrangement events, they share the same ancestral genomic context. As the context of a gene or genomic segment has functional consequences, toporthologous sequences are generally expected to be more similar in their function than orthologous sequences that are not toporthologous (atoporthologs) [6]. However, there is no guarantee that toporthologs share a common

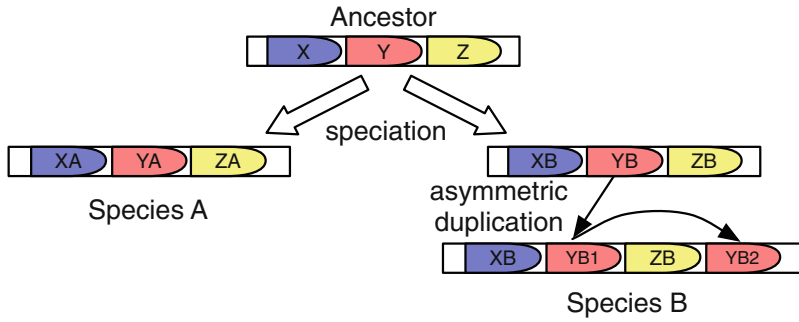

Fig. 1 A hypothetical evolutionary scenario in which the relation of toporthology distinguishes between two ortholog pairs. The bullet-like shapes indicate genomic segments. Both YB1 and YB2 are orthologous to YA. However, only YB1 is toporthologous to YA because YB2 was derived from the target of an asymmetric duplication since the time of the most recent common ancestor, Y, of YB2 and YA

function or that two genomic intervals that have the same function are toporthologs. Thus, a rigorous functional analysis of genomes should consider all classes of homology. Nevertheless, WGAs that focus on toporthology produce a good first approximation to a functional correspondence between genomes.

2.3 Definition and Representation

To be more precise, a WGA is, in general, the prediction of homologous pairs of positions between two or more genome sequences. Often, as we have previously discussed, only orthologous or toporthologous relations are predicted in WGAs. And while alignment is typically focused on homologous relationships *between* sequences, whole-genome comparisons can also include alignments *within* genomes, which represent paralogous sequences.

Note that we define WGA as homology prediction at the level of nucleotides. Although the concept of homology is more commonly used with respect to entire genes or proteins, it is easily used and, in fact, more naturally defined at the level of single nucleotides. Homology of nucleotide positions is established through template-driven nucleotide synthesis, and the definitions of orthology, paralogy, and xenology for nucleotides follow those for genes [7].

While a WGA can be defined as a prediction of homology statements, it is usually represented as a set of nucleotide-level alignment matrices or "blocks," each block made up by segments of the genomes that are both homologous and colinear. Homologous genomic segments are colinear if they have not been broken by a rearrangement event since the time of their MRCA. Since rearrangement events, such as inversions, are common at the scale of entire genomes, WGAs are typically made up of many blocks. In general, a block contains two or more genomic segments, and multiple segments in the same block may belong to the same genome (indicating paralogous sequence). One specific WGA representation, the "threaded blockset" [8], requires that every

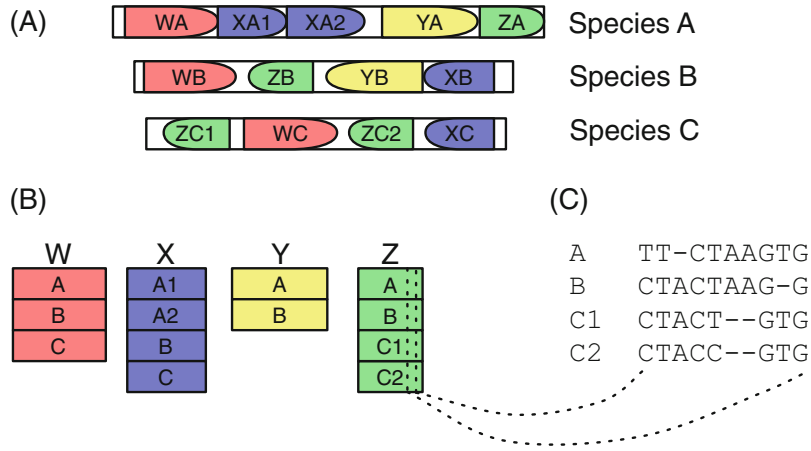

Fig. 2 An example WGA of three genomes represented as a set of alignment blocks. (**a**) The positions of the genomic segments that are in the alignment blocks are shown as shaded bullet-like shapes (the direction of the bullet indicates the orientation of the segment). In this example, not all genomic segments belong to a block (note the unshaded intervals). (**b**) The alignment blocks of the WGA. Note that blocks do not need to contain a segment from all genomes (e.g., block Y) and that some blocks can contain multiple segments from the same genome (e.g., blocks X and Z). (**c**) A slice of alignment block Z, which is a nucleotide-level alignment

position belongs to a block and thus additionally allows a block to contain just a single segment, which would represent a unique genomic sequence. Figure 2 depicts a hypothetical example of a WGA, with some blocks containing both orthologous and paralogous sequences.

As more genomes are added to an alignment or the total evolutionary divergence between them is increased, the blocks in a WGA decrease in size and increase in number. One might imagine that in the limit of an infinite number of genomes or an infinite amount of time, all blocks might have length one (a single column), which makes the concept of an "alignment matrix" irrelevant. However, rearrangements in certain segments of the genome are likely to be highly deleterious to an organism and will thus never be observed. Such segments are referred to as genomic "atoms" [9] and prevent all blocks from becoming single alignment columns.

2.4 Comparison to Other Homology Prediction Tasks

WGA is closely related to classical sequence alignment (the alignment of two or more relatively short and colinear sequences), and most whole-genome aligners rely on classical alignment techniques (e.g., the Needleman–Wunsch [10] and Smith–Waterman [11] pairwise alignment algorithms and heuristics used for multiple alignments) as subroutines. However, there are three key differences between these two classes of alignment. First, and most importantly, classical alignment requires sequences to be colinear, which is often not the case for genome sequences due to rearrangement events. Second, even when restricted to toporthologous relationships, the correspondences between genomes are not one to

one, which is also a requirement of classical alignment. Due, in part, to the complications of these first two issues, it is difficult to formulate a useful objective function (such as the sum-of-pairs score for classical alignment) for WGA. Thus, most genome alignment methods are heuristic procedures that lack an explicit objective. A last difference between classical alignment and WGA is the scale of the problem. Classical alignment typically focuses on the alignment of single genes, which are usually on the order of thousands of nucleotides long. Whole genomes, in contrast, are millions to billions of nucleotides in length. The facts that genomes are large and are often neither colinear nor in one-to-one correspondence with other genomes are what make WGA challenging.

Since WGA is often focused on orthologous relationships, it is also related to the "orthology prediction" problem (*see* Chapter 9 [5]). The key difference between the two problems is that orthology prediction is traditionally cast at the level of genes, whereas WGA operates at the level of nucleotides. For most orthology prediction methods, a genome is treated as an unordered set of genes. Whole-genome aligners, on the other hand, consider a genome to be a set of DNA sequences (chromosomes) within which genes are embedded. Thus, a WGA provides orthology predictions for both genes and intergenic regions. Due in part to their treatment of genomes as long nucleotide sequences, current WGA methods rely exclusively on sequence similarity and the ordering of nucleotides in a genome to predict orthology. In contrast, orthology prediction methods often use phylogenetic analyses, which can be more powerful than genome order and sequence similarity information alone. Thus, while the problem of WGA is broader in scope than that of orthology prediction, it is restricted to the analysis of relatively closely related genomes, for which homology of nongenic nucleotides is detectable and gene order is at least partially conserved. Gene-level orthology prediction is more appropriate for distantly related genomes, which may only have detectable homology at the amino acid level and little colinearity.

2.5 Significance

WGAs are powerful because they allow for the analysis of molecular evolution at both large and small scales. At the large scale, one can use such alignments to estimate the frequency and location of rearrangement and duplication events. For example, one might use a WGA between human and mouse to identify colinear orthologous blocks, which are then given to a rearrangement analysis method (e.g., [12]) to determine a most parsimonious set of rearrangement events explaining the current structures of the two genomes. At the small scale, WGAs can be used to examine the rates of substitutions and indels across the entire genome. For example, one might look at alignments of ancestral repeats to estimate the neutral rates of nucleotide evolution. Both small-

and large-scale mutational events identified from WGAs can be used as data for species tree inference. In combination with carefully constructed models of genome evolution at both scales, WGAs also enable the task of ancestral genome reconstruction [13, 14].

Beyond purely evolutionary studies, WGAs are valuable for identifying functional elements within genomes. Each class of functional element within the genome tends to have a unique "evolutionary signature," which can be searched for within WGAs [15]. For example, coding sequences tend to have mutational patterns with a predominance of substitutions at the third positions of codons, which are unlikely to affect the amino acid sequence. This characteristic evolutionary signature of coding sequence has led to the development of comparative gene-finding methods, which often use WGAs (Chapter 6 [16]). Noncoding RNA sequences can also be identified from WGAs but have more complex signatures involving compensatory mutations that maintain base pairing within RNA secondary structures [17]. More generally, one can search for evolutionarily constrained regions within WGAs, which can contain functional elements from a variety of classes [18]. When combined with the knowledge of transcription factor-binding motifs, this approach can be used to identify transcription factor-binding sites with a technique called "phylogenetic footprinting" [19]. The easiest evolutionarily constrained regions to pick out are those of "ultraconserved elements," which maintain high levels of sequence identity across large evolutionary distances and are primarily noncoding components of the genome [20].

WGAs also allow for the transfer of functional information about specific elements from one species to another. As WGAs typically predict orthology and orthologous sequences are likely to have similar functions, WGAs are valuable for function prediction. By aligning at the nucleotide level across the genome, they can aid in function prediction for both genes and nongenic regions, such as those that contain regulatory elements. For example, if we are interested in a specific disease-associated interval in the human genome, we might use an alignment to identify where its mouse orthologs are located. Knowledge of the mouse orthologs would enable us to have a better understanding of the evolutionary history of this genomic region and could lead to genetic manipulation experiments that can only be performed in mice.

3 Methods for WGA

3.1 A Simplistic Approach

It is easier to understand the existing methods for performing WGA by first appreciating the shortcomings of a simplistic approach for comparing whole-genome sequences. One simple approach would be to run BLAST [21], or another similar local alignment tool, between all pairs of genomes. The WGA would

then be defined as the union of all significant pairwise local alignments discovered by BLAST. By using a local alignment tool, we avoid the issues of rearrangements and duplications, as sets of local alignments are not constrained to be colinear or in one-to-one correspondence.

While this approach would certainly yield a large set of homology predictions between all pairs of genomes, it has a number of shortcomings. First, by only using a BLAST significance threshold, it makes no distinction between orthology, paralogy, and other refinements of homology. Second, the pairwise alignments that it produces are not guaranteed to be consistent with each other, even though homology, by definition, is a transitive relation. Third, BLAST may miss some homologous sequences that have low similarity but are strongly supported in their relatedness by flanking homologous sequences. BLAST's significance statistics are proven for ungapped sequences and good in practice for sequences with short indels [22], but are not designed for whole-genome comparisons, which often feature large-scale insertions and deletions and heterogeneous substitution rates. Lastly, this approach is overly computationally intensive. For example, it does not take advantage of the fact that homology is a transitive relation, that relationships between sequences are reasonably modeled by a tree, and that homologous sequences between genomes are often found in long colinear segments.

3.2 The Two Major Approaches to WGA

Existing WGA methods attempt to address one or more of the weaknesses of this simple approach. These methods can be loosely classified into two major strategies which we refer to as the "hierarchical" and "local" approaches. The main idea behind the hierarchical approach is to split the WGA problem into a set of global multiple alignment problems. To do this, it first identifies the colinear and homologous (typically orthologous) segments of the genomes. Each set of colinear segments is then given to a specialized genomic global alignment method to produce a nucleotide-level alignment. In contrast, the first step of the "local" approach is to produce a large set of nucleotide-level alignments. Later steps involve the filtering and merging of these alignments to produce sets of pairwise or multiple alignments of homologous (typically orthologous) sequences. Despite their differences, both strategies typically begin with a local alignment step that is similar to the simplistic all-vs.-all alignment of the BLAST approach. A summary of all of the WGA methods described in this chapter and the role they play within one or both approaches is given in Table 1.

Both approaches have advantages and disadvantages. The primary advantage of the hierarchical approach is that it can often be faster and breaks a WGA into a number of independent subproblems that can be solved in parallel. It is faster because the

Table 1
A list of the WGA methods cited in this chapter

Method	Category	Relationships predicted	Pairwise or multiple	References
BLAST	Local alignment	Homology	Pairwise	[21]
BLAT	Local alignment	Homology	Pairwise	[32]
STELLAR	Local alignment	Homology	Pairwise	[33]
LASTZ	Local alignment	Homology	Pairwise	[34]
LAST	Local alignment	Homology	Pairwise	[28]
MUMmer	Local alignment	Orthology	Pairwise	[35]
CHAOS	Local alignment	Homology	Pairwise	[36]
GRIMM-Synteny	Homology mapping	Toporthology	Multiple	[40]
DRIMM-Synteny	Homology mapping	Homology	Multiple	[45]
Mercator	Homology mapping	Toporthology	Multiple	[46]
Enredo	Homology mapping	Homology	Multiple	[47]
OSfinder	Homology mapping	Toporthology	Multiple	[48]
SuperMap	Homology mapping	Homology	Multiple	[49]
Sibelia	Homology mapping	Homology	Multiple	[50]
M-GCAT	Hierarchical WGA	Toporthology	Multiple	[51]
progressiveMauve	Hierarchical WGA	Toporthology	Multiple	[52]
MUGSY	Hierarchical WGA	Toporthology	Multiple	[53]
Cactus	Hierarchical WGA	Homology	Multiple	[54]
MAVID	Global genomic alignment	Colinear homology	Multiple	[60]
LAGAN/Multi-LAGAN	Global genomic alignment	Colinear homology	Pairwise/multiple	[37]
DIALIGN	Global genomic alignment	Colinear homology	Multiple	[36]
SeqAn::T-Coffee	Global genomic alignment	Colinear homology	Multiple	[61]
Pecan	Global genomic alignment	Colinear homology	Multiple	[47]
FSA	Global genomic alignment	Colinear homology	Multiple	[62]
NUCmer/PROmer	Local WGA	Orthology	Pairwise	[35]
MULTIZ/TBA	Local WGA	Homology	Multiple	[8]
AXTCHAIN/CHAINNET	Alignment chaining and filtering	Orthology	Pairwise	[67]

(continued)

Table 1
(continued)

Method	Category	Relationships predicted	Pairwise or multiple	References
PicoInversionMiner	Alignment refinement	Orthology	Pairwise	[68]
Cassis	Alignment refinement	Orthology	Pairwise	[69, 70]
GenAlignRefine	Alignment refinement	Colinear homology	Multiple	[71]
PSAR-Align	Alignment refinement	Colinear homology	Multiple	[73]
Phylo	Alignment refinement	Colinear homology	Multiple	[76, 77]
SLAM	Alignment refinement	Colinear homology	Pairwise	[78]
DOUBLESCAN	Alignment refinement	Colinear homology	Pairwise	[79]
CESAR	Alignment refinement	Colinear homology	Pairwise	[81]
MORPH	Alignment refinement	Colinear homology	Pairwise	[82]
EMMA	Alignment refinement	Colinear homology	Pairwise	[83]
MAFIA	Alignment refinement	Colinear homology	Multiple	[84]
SAPF	Alignment refinement	Colinear homology	Multiple	[85]
REAPR	Alignment refinement	Colinear homology	Multiple	[86]

For each method, the approach it uses or the role it plays within a larger WGA system is given in the "category" column. Each method is labeled as either "pairwise" or "multiple" depending on whether it can be applied to generate multiple alignments. In addition, the primary type of evolutionary relationship predicted by each method is given in the "relationships predicted" column

identification of long colinear and orthologous segments in the genomes can be accurately computed without the need for sensitive nucleotide-level alignments. However, because hierarchical methods do not often use the most sensitive aligners for this step, they tend to miss small rearranged or diverged segments. Thus, the primary advantage of the local method is in its sensitivity to these regions, although "glocal" alignment methods [23], which allow for small rearrangements, can partially ameliorate this weakness of hierarchical methods. Hierarchical methods also run the risk of being overconfident of the colinearity of genomic segments and can thus produce more false-positive aligned positions within sequences predicted to be colinear.

3.3 Local Pairwise Genomic Alignment

Methods for both WGA strategies generally start by finding local alignments between, and perhaps within, the genomes. The Smith–Waterman algorithm is the classical solution to the pairwise local alignment problem, but is generally not used for WGA because it runs in time quadratic in the size of the genomes, which can be large. Instead, most methods adopt a "seed-and-extend" approach for discovering high-scoring local alignments,

much like BLAST. This approach first identifies short ungapped matches between the sequences using one of a variety of data structures. It then extends the short matches from both ends using a variant of the Smith–Waterman algorithm, stopping the extension when the score of the alignment drops below a specified threshold. In some cases, nearby and consistent (in terms of order and orientation) local alignments are "chained" together to form larger alignments.

There are a number of techniques used for discovering seeds at the genomic scale for the "seed-and-extend" approach to local alignment. A first distinction between the techniques is whether they find exact or inexact matching seeds. Exact seed discovery is often faster and easier to implement, whereas inexact seeds offer better sensitivity. Seed techniques also vary in whether they use "consecutive" or "spaced" seeds [24]. Consecutive seeds consider matches and mismatches at all positions within a sequence interval, whereas spaced seeds only check for matches at a subset of positions within an interval. The specific subset of positions checked is known as the "seed pattern," and there has been significant work on determining optimal sets of multiple seed patterns (e.g., [25, 26]). It has been shown that carefully chosen spaced seed patterns are superior to consecutive seeds in terms of sensitivity [27]. Lastly, seeds differ in whether their lengths are fixed or adaptive (variable). For WGA, adaptive seeds have been shown to allow for faster local alignment at the same level of sensitivity as fixed seeds [28].

Seed-finding techniques can often be improved by taking advantage of DNA evolutionary models. A generalization of spaced seeds is "subset seeds" [29], which allow subsets of bases to be considered equivalent when determining if there is a match at a given position. Subset seeds are particularly useful for taking into account that transitions are often more common than transversions in genome comparisons. Further taking into account biologically informed substitution patterns is the "translated" seed, which is a match at the amino acid level after translating genomic sequences in all six possible reading frames. Translated seeds enable increased sensitivity in comparisons of more diverged genomes. Lastly, when aligning a genome to a set of genomes for which a multiple WGA has already been constructed, one can take into account the substitution patterns and ancestral sequences inferred from the WGA to devise more sensitive seeds [30, 31].

The choice of seed type is the major determinant of the data structures used for seed discovery. For example, BLAT [32] uses a simple index of all possible k-mers for exact and translated seeds but uses a heuristic of indexing only nonoverlapping k-mers for memory efficiency. STELLAR [33] also uses an index of k-mers but implements an exact algorithm based on filtration for finding all local alignments with an error rate below a given threshold. LASTZ

(the successor to BLASTZ [34]), which uses a carefully chosen spaced seed pattern introduced by [24], instead uses a hash table to find both exact and inexact matches. Not to be confused with LASTZ is the more recently developed LAST aligner [28], which uses adaptive seeds with highly configurable patterns that are identified via a suffix array data structure. MUMmer uses a suffix tree to rapidly find all exact consecutive seeds with some minimum length [35]. CHAOS [36], which is a component of the LAGAN-suite of genome alignment tools [37], uses a related structure, a "threaded trie," to find exact and inexact consecutive seeds.

For computational efficiency reasons, the extension step of the seed-and-extend approach typically only allows for ungapped alignments or alignments with short indels. However, genome alignments often feature large indels that are not discovered by extension from a seed. Thus, many local genomic alignment tools use a "chaining" step to link nearby and consistent local alignments discovered by the seed-and-extend strategy. For example, MUMmer includes a module for chaining together nearby exact matches using a variation of the longest increasing subsequence (LIS) problem [38]. CHAOS also uses an LIS-derived algorithm for chaining the inexact consecutive seeds it discovers. Chaining is often followed by more sensitive alignment between chained local alignments. For example, MUMmer runs a variant of Smith–Waterman alignment in between chained matches and LASTZ recursively searches for alignments with more sensitive seeds in between nearby alignments discovered in previous steps.

3.4 The Hierarchical Approach

The hierarchical approach to WGA consists of two steps. First, a high-level homology map between the genomes is constructed. Second, a nucleotide-level alignment is obtained by running a genomic global alignment tool on each homologous and colinear set of genomic segments identified by the homology map. Hierarchical WGA methods vary in the exact techniques used for each step.

The idea behind the hierarchical approach is to separate the problem of identifying rearrangements and duplications from that of obtaining a nucleotide-level alignment. In the absence of rearrangements and duplications, WGA simply reduces to classical sequence alignment although at a much larger scale. Thus, if a WGA problem can be broken into a set of subproblems that do not contain these large-scale events, the numerous methods that have been developed for classical global alignment can be utilized.

The first step of the hierarchical strategy is to construct a homology map between the genomes of interest. A homology map is a collection of sets of genomic intervals, where each set of intervals is required to be homologous and colinear (i.e., free of rearrangements and duplications). Each set represents the sequences that will ultimately form a block within a WGA.

Homology maps generally have the property that each genomic position belongs to at most one set and has all of its homologs contained within that set. For WGA, homology maps are often restricted in the evolutionary relationships that are captured, as only a subset of homologous relationships may be of interest. Typically, only orthologous relationships are captured, forming an "orthology map." When orthology maps are restricted to predicting one-to-one relationships, they are more likely to be representative of toporthology.

The concept of a homology map is closely related to the concepts of "conserved segments" and "syntenic blocks," which generally refer to sets of genomic intervals containing multiple homologous markers (e.g., genes) and featuring conserved orientations and adjacencies of these markers [39, 40]. Unfortunately, these concepts have long been poorly defined, and, as a result, methods for syntenic block identification differ markedly in their output [41]. In addition, methods for identifying syntenic blocks (or closely related concepts) are often focused on identifying sets of genomic intervals that exhibit levels of conservation of marker content or colinearity that exceed what one would expect if markers were randomly shuffled between genomes (e.g., [42–44]). This is in contrast to homology maps, which are concerned with colinear homology, regardless of biological significance. And, in practice, homology maps are intermediate objects in the process of WGA, whereas syntenic block predictions are often of direct interest.

Homology maps are most commonly constructed from local alignments, such as those computed by methods discussed in the previous section. As only a high-level correspondence is desired, these methods are often run in faster but less sensitive configurations. For example, local alignments between just the coding intervals of the genomes can be computed quickly and used for the construction of homology maps that are at least accurate with respect to protein-coding genes.

Although numerous pairwise homology mapping methods exist, in this chapter, we restrict our attention to methods that scale to more than two genomes, as the problem is significantly more challenging in the multiple genome case. Examples of multiple genome homology map methods include GRIMM-Synteny [40], its successor DRIMM-Synteny [45], Mercator [46], Enredo [47], OSfinder [48], SuperMap [49], and Sibelia [50]. The WGA programs M-GCAT [51], progressiveMauve [52], MUGSY [53], and Cactus [54] are integrated hierarchical methods that contain a homology mapping stage.

Many of these methods use graph-based data structures to find a mapping between multiple genomes simultaneously. Kehr et al. [55] characterized the relationships between four commonly used types of graphs: alignment graphs [56], *A-Bruijn* graphs [57, 58], Enredo graphs [47], and Cactus graphs [59]. The most

straightforward graph is the alignment graph, which is a mixed graph with vertices representing genomic segments, directed edges representing adjacent segments, and undirected edges representing homologous segments. In an *A-Bruijn* graph, vertices instead represent sets of homologous segments, and directed edges represent adjacencies between pairs of segments (one from each set represented by the connected vertices). Relative to alignment graphs, *A-Bruijn* graphs are more compact and readily reveal the content of each genome. An Enredo graph is very similar to an *A-Bruijn* graph, but has a pair of vertices instead of a single vertex for each set of homologous segments, which captures information regarding the directionality of each segment within a homologous set. Lastly, cactus graphs flip the representation of adjacencies, with vertices corresponding to sets of adjacencies and edges corresponding to sets of homologous segments. Cactus graphs have a natural decomposition that provides advantages for analysis and visualization of WGAs.

Graph-based homology mapping methods generally produce an initial WGA graph using one of the four representations we have discussed and then refine the graph via modifications. Of the homology mapping methods we have listed, GRIMM-Synteny, Mercator, and MUGSY use alignment graphs. DRIMM-Synteny and OSfinder use *A-Bruijn* graphs and Sibelia uses de Bruijn graphs, of which *A-Bruijn* graphs are a generalization. And, as their names suggest, Enredo and Cactus use Enredo and cactus graphs, respectively. These methods use a variety of techniques for graph refinement. For example, MUGSY is unique in its use of flow network algorithms to identify breaks in colinearity. OSfinder uses a novel probabilistic model to determine a maximum likelihood multiple genome orthology map. And Cactus uses a simulated annealing-style algorithm, the *Cactus alignment filter*, to refine an initial cactus graph representing a homology map.

Unlike the graph-based methods that build a map between all genomes simultaneously, the SuperMap and progressiveMauve methods build a multiple genome map by progressively building pairwise maps up a guide tree. The pairwise SuperMap algorithm is essentially a symmetric version of the chaining method used by Shuffle-LAGAN [23], which allows for rearrangements and duplications in its chains of orthologous segments. The progressiveMauve mapping method instead uses a "breakpoint elimination" algorithm to find colinear segments and does not allow for duplications, thus producing output indicative of one-to-one toporthology. This algorithm greedily removes local alignments one by one with the goal of maximizing an objective function that takes into account both the number of breakpoints implied by an alignment and substitution scores.

Once a homology map has been created, any one of a number of genomic global alignment methods can be used to align the orthologous and colinear segments identified by the map. As for

our discussion of homology mapping methods, we restrict our attention to global aligners that can handle multiple genomes. Examples of such methods are MAVID [60], MLAGAN [37], DIALIGN [36], SeqAn::T-Coffee [61], PECAN [47], FSA [62], and the *base-level alignment refinement* (*BAR*) algorithm of Cactus [54]. For colinear sequences, the genomic alignment problem is the same as that of classical global alignment but is made more difficult by the fact that the sequences are long (possibly millions of nucleotides in length). Thus, global genomic aligners employ heuristics to speed up the process. By far, the most common heuristic used is to first identify short local alignments, or *anchors*, between the sequences, identify a chain of these anchors, and then perform global alignment between the adjacent chained anchors. This technique is similar to the strategy for hierarchical WGA, but is simpler, due to the fact that rearrangements and duplications do not need to be taken into account. MLAGAN and DIALIGN use the CHAOS local aligner, PECAN and FSA use Exonerate [63], and MAVID and SeqAn::T-Coffee use suffix trees or arrays to find anchors.

In addition to the specific local alignment technique used to speed up the alignment process, global genomic aligners also vary with respect to how they combine local pairwise alignments to build a multiple global alignment. First, MAVID, MLAGAN, SeqAn::T-Coffee, and Pecan all belong to the class of progressive alignment methods, which use a phylogenetic tree to guide their algorithms (*see* Chapter 7 [1]). For the alignment of non-leaf sequences during progressive alignment, MAVID uses maximum likelihood ancestral sequence inference, while MLAGAN, SeqAn::T-Coffee, and Pecan use a sum-of-pairs objective function. Both SeqAn::T-Coffee and Pecan use a "consistency" technique, which adjusts the score between pairs of positions (or segments) based on the consistency of triplets of pairwise alignments. The nonprogressive methods, DIALIGN, FSA, and BAR, instead put together a multiple alignment by greedily merging consistent local pairwise alignments. While differing in their use of a tree, the FSA, Pecan, and BAR methods take advantage of probabilistic models of sequence alignment and attempt to maximize statistically grounded objective functions, as opposed to the heuristic score-based functions used by the other methods. BAR is unique in its ability to predict breakpoints when aligning groups of sequences that may contain the boundaries of rearrangement events.

Although the hierarchical approach breaks the WGA problem into a large number of subproblems (one per colinear segment set) that can be computed in parallel, it is still a significant computational effort to produce a WGA with this approach, particularly for large eukaryotic genomes. Thus, a number of Web sites host precomputed hierarchical WGAs. Alignments produced by the combination of Pecan with either Enredo or Mercator are hosted at the Ensembl Web site [64]. Similarly, the VISTA Web site [65] hosts

WGAs generated by SuperMap and the LAGAN-suite of genomic aligners. Both sites offer visualizations of the WGAs, which are useful for looking at levels of conservation across genomes.

3.5 The Local Approach

The local approach to WGA bypasses the high-level homology map construction phase of the hierarchical approach and instead begins by identifying a comprehensive set of nucleotide-level pairwise local alignments. The second step of this approach is to combine the pairwise local alignments into a cohesive WGA by filtering out nonorthologous relationships and merging pairwise alignments into multiple alignments. Because there is typically no additional pairwise nucleotide-level alignment performed in the second step, the local alignments generated by the first step are obtained with a more sensitive aligner than that used by hierarchical methods for homology map building. The two primary examples of local WGA methods are MUMmer, a pairwise genome aligner, and MULTIZ/TBA, a multiple genome aligner [8].

MUMmer was one of the first pairwise WGA methods to be developed and was initially targeted at the alignment of prokaryotic-sized genomes. The WGA ability of MUMmer is achieved through a combination of smaller modules that is orchestrated by the NUCmer or PROmer scripts. The first module identifies maximum unique matches (MUMs) between a pair of genomes with a suffix tree data structure. Nearby matches are clustered together, and a high-scoring colinear chain of matches is identified within each cluster. Finally, the matches within the chains are extended with a variant of the Smith–Waterman algorithm, and the resulting extended chains are output as a WGA. The raw WGA output by MUMmer can, in general, include all classes of homologous relationships. However, the chains are typically filtered to leave only those that are highest scoring or that result in a reference position being overlapped by only a single chain. Thus, a filtered WGA from MUMmer is usually representative of orthology.

MULTIZ/TBA, which was instead designed for large eukaryotic genomes, starts by using LASTZ to generate sensitive local pairwise alignments between all pairs of genomes or between a reference genome and all others. MULTIZ is then used to identify local alignment blocks of subsets of genomes that should be combined and to merge these blocks using a banded variant of the Smith–Waterman algorithm. TBA is the program that is used to coordinate this entire process when all pairs of genomes are compared. Thus far, it does not appear that TBA has been used at the whole-genome scale, although MULTIZ is regularly used for reference-based WGAs hosted by the UCSC Genome Browser [66]. For these reference-based WGAs, the ungapped segments of LASTZ alignments are first processed with a chaining program (AXTCHAIN) to establish large colinear alignments between the reference and another genome. In contrast to the output of

chaining methods discussed in Subheading 3.3, a chain produced by AXTCHAIN is an ordered set of pairwise local alignments rather than a single long alignment that explicitly aligns between the short local alignments that form the chain. AXTCHAIN chains are typically filtered by the CHAINNET program to retain only the highest-scoring alignment at each position within the reference genome [67]. The remaining alignments, which most likely reflect orthologous relationships, are then combined into multiple alignments with MULTIZ.

3.6 Refining WGAs

Because of the computational complexity of multiple alignment, particularly at the whole-genome scale, methods of both approaches to WGA use heuristics and simplified models to make WGA feasible. For example, most of the methods described in this chapter do not distinguish between different classes of genomic sequence (e.g., genic and intergenic) while constructing nucleotide-level alignments. And many methods disregard small, marginally significant, local alignments for the sake of speed. As a result, at a local level, the results of current WGA methods often leave room for improvement.

To remedy this situation, a number of methods have been developed that may be used to refine WGAs. These methods take as input either a WGA, a single WGA block, or the set of homologous and colinear sequences that make up a WGA block. They can be generally grouped into one of three categories. The first is composed of methods that refine the local structure of a WGA. That is, they redefine the boundaries, or "breakpoints," of the homologous and colinear blocks in the WGA. A secondary category of methods focuses on optimizing individual WGA blocks with respect to an objective function. The last category includes methods that perform alignment while taking into account the structure and evolutionary dynamics of certain classes of genomic elements.

PicoInversionMiner [68] and Cassis [69, 70] are two methods for refining the local structure of a WGA. PicoInversionMiner identifies very small "inplace" inversions between two genomes that are left undetected by an initial WGA. Such inversions are represented by alignments that would typically not have statistically significant scores at the genome level but can be detected via probabilistic models of local sequence evolution. In contrast to PicoInversionMiner, which identifies novel rearrangement events, Cassis refines the coordinates of breakpoints. The refinements produced by Cassis are the result of identifying weak similarities between sequences adjacent to segments of an initial orthology map and extending the boundaries of segments based on these similarities. The BAR algorithm of Cactus, which we have previously discussed in the context of hierarchical WGA, is also an alignment refinement method that identifies breakpoints.

Other methods for refining WGAs focus on improving local colinear multiple alignments with respect to a given objective function. For example, GenAlignRefine [71] attempts to optimize WGA blocks according to the COFFEE objective function [72] using a genetic algorithm. The PSAR-Align method [73] instead realigns blocks to optimize an expected accuracy objective function [74] using pairwise alignment probabilities estimated by the PSAR tool [75] and the sequencing annealing algorithm of the FSA multiple alignment method [62]. Lastly, the Phylo project [76, 77] refines WGAs by "crowd sourcing" the task of optimizing colinear alignment blocks, according to one of a number of objective functions. Phylo casts the multiple alignment problem as a casual game that may be played by "citizen scientists" at the project's website (http://phylo.cs.mcgill.ca/).

Lastly, a number of methods have been developed that can improve the alignments of specific classes of genomic elements, such as gene structures. The primary goal of these methods is generally to improve prediction of genomic elements, but a more accurate alignment often results as a side product. Among the oldest of such methods are comparative gene finders that perform protein-coding gene prediction and pairwise alignment simultaneously. These include SLAM [78] and DOUBLESCAN [79], both of which use pair hidden Markov models [80]. A related method, CESAR [81], was specifically designed for realignment and targets individual coding exons rather than full gene structures. Other methods focus on improving the alignment of noncoding regulatory regions by modeling the evolution of sets of transcription factor-binding sites with known motifs (e.g., MORPH [82], EMMA [83], and MAFIA [84]). Like the comparative gene finders, these methods also use statistical alignment techniques but with models extended to take into the account the conservation of binding sites instead of gene structures. SAPF [85] is also a method aimed at alignment of noncoding regulatory regions but more generally models sequences that are mixtures of "slow" and "fast" evolving elements without knowledge of binding motifs. Lastly, REAPR [86] focuses on the realignment and detection of noncoding RNAs by using alignment models that take into account the conserved secondary structures of such RNAs.

4 Evaluation of WGAs

Just as for small-scale alignment (Chapter 7, [1]), assessing the accuracy of WGAs is hard because we rarely know the true evolutionary history of a set of genome sequences. In fact, the evaluation of WGAs is even harder than that of protein alignments. While protein aligners can be evaluated with "gold standard" benchmarking databases where the truth is established through protein

structural information, genome aligners have no benchmarks of real data. In addition, WGAs must be assessed not only for whether they align truly homologous sequences but also for whether they correctly predict orthologous (or toporthologous) relationships. Thus, the evaluation of WGAs is related to that of gene orthology prediction, which is discussed in Chapter 9 [5]. Despite these challenges, a number of creative approaches have been used for determining the accuracy of WGA methods. The approaches generally fall into four categories: (1) simulation, (2) analysis of alignments to annotated regions, (3) comparison with predictions from other methods, and (4) alignment statistics.

Simulated data are appealing for evaluation as we know the entire evolutionary history of the simulated sequences and can thus thoroughly evaluate the accuracy of an alignment. Many of the WGA methods described in this chapter have used simulations for assessing their accuracies [8, 47, 52, 54, 62]. The Alignathon [87], one of the most comprehensive evaluations of WGA methods to date, relied heavily on simulated data sets. This study called attention to one potential pitfall of simulation-based evaluation, which is that the performance of a WGA method may be overestimated when that method was developed or trained with respect to the same simulator used for the assessment.

Simulating the evolution of whole genomes is a challenging task, and it is unclear if the current models used for simulation are close to reality. Such models are highly complex, as they have to account for many different types of evolutionary events, at both the small and large scales. For example, they need to model the random mutations of both single-nucleotide substitutions and megabase-sized inversions. In addition, they also need to model natural selection, which alters the probability of these random mutations becoming fixed within a population. For example, an inversion that cuts an essential gene in half might have a much lower probability of becoming fixed than an inversion with both end points in intergenic regions. Despite these challenging model details, a number of genomic evolution simulators have been developed. Currently, only three simulators model both small-scale events (e.g., substitutions and indels) and large-scale rearrangements and duplications [88–90]. Other simulators focus only on nonrearranging events [8, 91–98] and are thus good for evaluating colinear genomic aligners but not homology mapping methods.

A second class of approaches to evaluating WGAs leverages our knowledge of various classes of elements within the genome. For example, with our understanding that most coding regions are conserved across closely related genomes, the fraction of exons in a genome "covered" by an alignment is an indirect measure of the sensitivity of a WGA [37, 49, 60, 99]. Specificity can also be roughly assessed with coding regions, either by counting the number of coding bases that are aligned to noncoding bases in other

genomes [36, 100] or by checking that alignments in coding regions exhibit periodicities in their substitution patterns [99]. A related approach that instead assesses the accuracy of eukaryotic orthology maps is to check if exons from the same gene are mapped in the same order and orientation to other genomes [47]. For the subset of protein-coding and noncoding RNA genes that have curated "gold standard" alignments, the accuracy of a WGA with respect to those genes may be assessed [101]. However, the fact that genic regions are often highly conserved is also a disadvantage of using them for evaluation; the most conserved regions are the easiest to align, and some aligners use exon annotation information or translated matches. Because of these issues, repeat sequences, which are believed to evolve more neutrally, have been used for alignment evaluation [47, 99]. For example, in [99], sensitivity was assessed by alignments of ancestral repetitive elements, and specificity was inferred from the number of alignments to lineage-specific repeat elements (in this study, primate-specific *Alu* repeats).

Another common evaluation technique is to compare whole-genome aligners against other related methods. For example, a WGA produced by one method can be used as the "truth" with which to evaluate the sensitivity and specificity of other WGAs [53]. This technique is useful for judging the similarity of different WGAs but, unfortunately, does not provide much information about accuracy. Another technique is to compare with the results from gene orthology prediction programs [48, 49]. The advantage of this approach is that it provides a more independent test of accuracy, since gene orthology prediction programs generally use different algorithms and information sources to infer orthology. The disadvantages of this approach are that it only provides a gene-level measure of accuracy and does not evaluate alignments of noncoding regions. In addition, since WGA and gene orthology prediction share similar goals, we might expect that future methods will blend techniques from both and thus that this evaluation approach will decrease in usefulness.

A last class of evaluation techniques involves the computation of statistics for WGAs. These statistics can be subdivided into simple descriptive statistics and measures computed via statistical or sampling techniques. One of the most straightforward descriptive statistics of a WGA is the "coverage" or the fraction of the genomes included in an alignment or orthology map block [45, 47, 49, 53, 87]. Generally, the higher the coverage, the more sensitive the WGA is believed to be, although one can easily create high-coverage WGAs with poor sensitivity. As a check of large-scale specificity in mammalian WGAs, the authors of [47] checked the fraction of the X chromosome that was covered by alignments to autosomal chromosomes in other genomes (the assumption being that translocations into and out of the X chromosome are rare in mammals). Some more detailed nucleotide-level statistics of WGAs

include the total number of "core" positions [53], which are gap-free alignment columns containing all genomes, and the average level of sequence identity in aligned columns [61].

More sophisticated statistics related to WGA accuracy are computed through the use of statistical or sampling techniques. Just as they are used for BLAST, Karlin and Altschul statistics [102] may be used to assess the significance of local pairwise alignments between genomes. StatSigMA extends these statistics to multiple alignments [103], and StatSigMA-w further extends this technique to detect dubiously aligned regions in WGAs of multiple genomes [104]. Whereas a given local pairwise alignment may be highly significant, the flanks of that alignment may be spurious, and a *p*-value may be computed assessing the possible "over-alignment" of a flank [105]. Within a multiple alignment, a number of techniques have been developed for estimating the accuracy of the alignment of pairs of residues or entire columns, including simply computing an alignment of reversed sequences [106], computing alignments with bootstrapped guide trees [107], sampling suboptimal multiple alignments [75], and evaluating consistency within a library of alternative alignments [108].

5 Future Challenges

Despite the substantial progress made in WGA methodology development, there are a number of challenges that remain unsolved. First, we are in need of WGA methods that can scale to hundreds or thousands of genomes. Along with ever-improving sequencing technology, we are accumulating whole-genome sequences at an increasing rate. Projects such as the Genome 10K Community of Scientists [109], which aims to collect and sequence the genomes of 10,000 vertebrate species, will further push the WGA problem to new scales. While most WGA algorithms have been made efficient for long genomes, very few are practical for large numbers of genomes. Encouragingly, we are beginning to see methods capable of scaling to thousands of genomes for the simpler task of "core-genome alignment" of highly similar microbial-sized genomes [110]. However, methods scaling to thousands of genomes for the full WGA task or for mammalian-sized genomes do not currently exist. In addition to algorithmic advances, we will also be in need of novel approaches for storing and representing WGAs of thousands of genomes.

Second, advances are needed in the parameterization of WGA methods. Current methods are littered with large numbers of parameters that are often heuristic in nature and not easily determined. In some cases, the default parameters for a WGA method may be markedly suboptimal [111]. One solution to this problem is to adopt probabilistic models, which offer principled approaches to

parameter estimation, such as maximum likelihood. In fact, probabilistic models of sequence evolution have already been adopted for the alignment of colinear genomic segments and have been shown to offer improved accuracy [47, 62]. However, we have yet to see a method that integrates probabilistic models of both small- and large-scale changes that is capable of constructing an entire WGA, although the recently introduced "split-alignment" pairwise WGA method is a promising step in this direction [112]. In addition, most WGA alignments use models or scoring schemes that assume homogenous rates of evolution across the genome. This assumption is obviously violated in real data, and new methods will need to be developed that take this into account. Simulated noncoding genomic alignments that represent a heterogeneous mix of evolutionary rates have been developed and should be useful for the development of new WGA methodology [97].

Lastly, more attention must be paid to the fact that a WGA is typically just a single estimate of the evolutionary history of a set of genomes and portions of this estimate may be highly uncertain. Encouragingly, methods for colinear genomic alignment have brought light to this issue at the nucleotide level [62, 113]. However, the issue of uncertainty at the large-scale orthology map level has not been sufficiently studied, perhaps due to the lack of probabilistic models for that level of the WGA problem. In addition, most efforts to address uncertainty in alignments simply assign levels of confidence to the components of a single alignment. It may be more useful to be presented with a set of near-optimal alignments so that alternative evolutionary histories can be examined by downstream analyses [114]. The determination and representation of uncertainty for all scales of a WGA will likely remain a challenging problem as the number of genomes included in alignments increases.

6 Exercises

1. Download the whole-genome aligner MUMmer (http://mummer.sourceforge.net) and FASTA-formatted genome sequences for the species *Helicobacter pylori J99* and *Helicobacter pylori B38* from GenBank (http://www.ncbi.nlm.nih.gov/genbank/, accessions NC000921 and NC012973, respectively). Run the NUCmer or PROmer programs on the two genome sequences. Visualize the resulting alignment with the mummerplot program. How many colinear blocks are there in the alignment? How many inversion events are implied by the alignment?

2. Visit the UCSC Genome Browser (http://genome.ucsc.edu) and browse the human genome version GRCh38/hg38. Search for and view the *CFTR* gene, mutations in which cause the disease cystic fibrosis. Turn on the Net tracks for alignments to

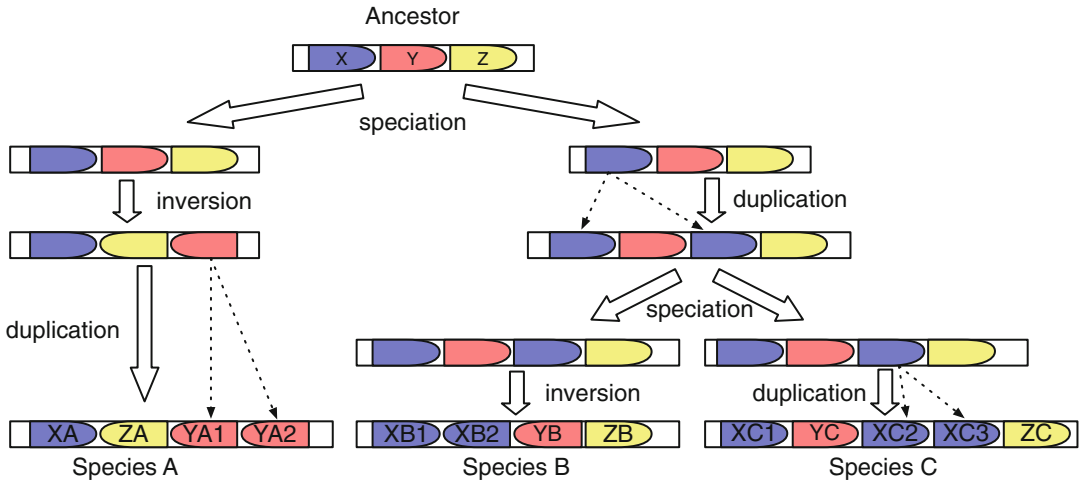

Fig. 3 The evolutionary scenario to be considered for Exercise 3. Each bullet-like shape corresponds to a genomic segment, with the direction of the bullet indicating the orientation of the segment

genomes of non-primate placental mammals by clicking on the "Placental Chain/Net" link (in the "Comparative Genomics" section) and choosing the appropriate configuration. Examine the Mouse Net track in the visualization and note the color of the mouse net alignments. Using the "Chromosome Color Key" (located in between the browser visualization and the track configuration section), identify the chromosome on which the mouse ortholog of *CFTR* is located. Looking at the net alignments for all of the placental mammals, does it appear that *CFTR* has been conserved across this clade?

3. Consider the evolutionary scenario giving rise to the genomes of three species shown in Fig. 3. For each of the relations listed below, give the pairs of genomic segments with that relation.

(a) Orthology

(b) Paralogy

(c) Toporthology

References

1. Löytynoja A (2012) Alignment methods: strategies, challenges, benchmarking, and comparative overview. Methods Mol Biol 855:203–235
2. Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512
3. Mukherjee S, Stamatis D, Bertsch J et al (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res 45:D446–D456
4. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113
5. Altenhoff AM, Dessimoz C (2012) Inferring orthology and paralogy. Methods Mol Biol 855:259–279
6. Dewey CN (2011) Positional orthology: putting genomic evolutionary relationships into context. Brief Bioinform 12(5):401–412

7. Dewey CN, Pachter L (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. Hum Mol Genet 15 Spec No 1:R51–R56
8. Blanchette M, Kent WJ, Riemer C et al (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708–715
9. Ma J, Ratan A, Raney BJ et al (2008) The infinite sites model of genome evolution. Proc Natl Acad Sci U S A 105:14254–14261
10. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
11. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
12. Tesler G (2002) GRIMM: genome rearrangements web server. Bioinformatics 18:492–493
13. Paten B, Herrero J, Fitzgerald S et al (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 18:1829–1843
14. Ma J, Zhang L, Suh BB et al (2006) Reconstructing contiguous regions of an ancestral genome. Genome Res 16:1557–1565
15. Stark A, Lin MF, Kheradpour P et al (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450:219–232
16. Alioto T (2012) Gene prediction. Methods Mol Biol 855:175–201
17. Eddy SR (2002) Computational genomics of noncoding RNA genes. Cell 109:137–140
18. Margulies EH, Blanchette M, NISC Comparative Sequencing Program et al (2003) Identification and characterization of multi-species conserved sequences. Genome Res 13:2507–2518
19. Tagle DA, Koop BF, Goodman M et al (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203:439–455
20. Bejerano G, Pheasant M, Makunin I et al (2004) Ultraconserved elements in the human genome. Science 304:1321–1325
21. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
22. Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
23. Brudno M, Malde S, Poliakov A et al (2003) Glocal alignment: finding rearrangements during alignment. Bioinformatics 19(Suppl 1):i54–i62
24. Ma B, Tromp J, Li M (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics 18:440–445
25. Sun Y, Buhler J (2005) Designing multiple simultaneous seeds for DNA similarity search. J Comput Biol 12:847–861
26. Xu J, Brown D, Li M et al (2006) Optimizing multiple spaced seeds for homology search. J Comput Biol 13:1355–1368
27. Zhang L (2007) Superiority of spaced seeds for homology search. IEEE/ACM Trans Comput Biol Bioinform 4:496–505
28. Kiełbasa SM, Wan R, Sato K et al (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21:487–493
29. Kucherov G, Noé L, Roytberg M (2006) A unifying framework for seed sensitivity and its application to subset seeds. J Bioinform Comput Biol 4:553–569
30. Flannick J, Batzoglou S (2005) Using multiple alignments to improve seeded local alignment algorithms. Nucleic Acids Res 33:4563–4577
31. Sun H, Buhler JD (2012) PhyLAT: a phylogenetic local alignment tool. Bioinformatics 28:1336–1344
32. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664
33. Kehr B, Weese D, Reinert K (2011) STELLAR: fast and exact local alignments. BMC Bioinform 12:S15
34. Schwartz S, Kent WJ, Smit A et al (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103–107
35. Delcher AL, Kasif S, Fleischmann RD et al (1999) Alignment of whole genomes. Nucleic Acids Res 27:2369–2376
36. Brudno M, Chapman M, Göttgens B et al (2003) Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinform 4:66
37. Brudno M, Do CB, Cooper GM et al (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731
38. Gusfield D (1997) Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press, Cambridge

39. Nadeau JH, Taylor BA (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci U S A 81:814–818
40. Pevzner P, Tesler G (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res 13:37–45
41. Ghiurcuta CG, Moret BME (2014) Evaluating synteny for improved comparative studies. Bioinformatics 30:i9–i18
42. Wang X, Shi X, Li Z et al (2006) Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. BMC Bioinform 7:447
43. Proost S, Fostier J, De Witte D et al (2012) i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res 40:e11
44. Lucas JMEX, Muffato M, Roest Crollius H (2014) PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. BMC Bioinform 15:268
45. Pham SK, Pevzner PA (2010) DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. Bioinformatics 26:2509–2516
46. Dewey CN (2007) Aligning multiple whole genomes with Mercator and MAVID. In: Bergman N (ed) Methods in Molecular Biology, vol 395. Humana, Clifton, NJ, pp 221–236
47. Paten B, Herrero J, Beal K et al (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res 18:1814–1828
48. Hachiya T, Osana Y, Popendorf K et al (2009) Accurate identification of orthologous segments among multiple genomes. Bioinformatics 25:853–860
49. Dubchak I, Poliakov A, Kislyuk A et al (2009) Multiple whole-genome alignments without a reference organism. Genome Res 19:682–689
50. Minkin I, Patel A, Kolmogorov M et al (2013) Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In: Algorithms in bioinformatics, Lecture notes in computer science. Springer, Berlin, pp 215–229
51. Treangen TJ, Messeguer X (2006) M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. BMC Bioinform 7:433
52. Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147
53. Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342
54. Paten B, Earl D, Nguyen N et al (2011) Cactus: algorithms for genome multiple sequence alignment. Genome Res 21:1512–1528
55. Kehr B, Trappe K, Holtgrewe M et al (2014) Genome alignment with graph data structures: a comparison. BMC Bioinform 15:99
56. Kececioglu J (1993) The maximum weight trace problem in multiple sequence alignment. In: Combinatorial pattern matching, Lecture notes in computer science. Springer, Berlin, pp 106–119
57. Pevzner PA, Pevzner PA, Tang H et al (2004) De novo repeat classification and fragment assembly. Genome Res 14:1786–1796
58. Raphael B, Zhi D, Tang H et al (2004) A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res 14:2336–2346
59. Paten B, Diekhans M, Earl D et al (2011) Cactus graphs for genome comparisons. J Comput Biol 18:469–481
60. Bray N, Pachter L (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res 14:693–699
61. Rausch T, Emde AK, Weese D et al (2008) Segment-based multiple sequence alignment. Bioinformatics 24:i187–i192
62. Bradley RK, Roberts A, Smoot M et al (2009) Fast statistical alignment. PLoS Comput Biol 5:e1000392
63. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinform 6:31
64. Herrero J, Muffato M, Beal K et al (2016) Ensembl comparative genomics resources. Database (Oxford) 2016:bav096
65. Brudno M, Poliakov A, Minovitsky S et al (2007) Multiple whole genome alignments and novel biomedical applications at the VISTA portal. Nucleic Acids Res 35: W669–W674
66. Casper J, Zweig AS, Villarreal C et al (2018) The UCSC Genome Browser database: 2018 update. Nucleic Acids Res 46:D762–D769
67. Kent WJ, Baertsch R, Hinrichs A et al (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100:11484–11489

68. Hou M, Yao P, Antonou A et al (2011) Picoinplace-inversions between human and chimpanzee. Bioinformatics 27:3266–3275
69. Lemaitre C, Tannier E, Gautier C et al (2008) Precise detection of rearrangement breakpoints in mammalian chromosomes. BMC Bioinform 9:286
70. Baudet C, Lemaitre C, Dias Z et al (2010) Cassis: detection of genomic rearrangement breakpoints. Bioinformatics 26:1897–1898
71. Wang C, Lefkowitz EJ (2005) Genomic multiple sequence alignments: refinement using a genetic algorithm. BMC Bioinform 6:200
72. Notredame C, Holm L, Higgins DG (1998) COFFEE: an objective function for multiple sequence alignments. Bioinformatics 14:407–422
73. Kim J, Ma J (2014) PSAR-align: improving multiple sequence alignment using probabilistic sampling. Bioinformatics 30:1010–1012
74. Schwartz AS, Pachter L (2007) Multiple alignment by sequence annealing. Bioinformatics 23:e24–e29
75. Kim J, Ma J (2011) PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucleic Acids Res 39:6359–6368
76. Kawrykow A, Roumanis G, Kam A et al (2012) Phylo: a citizen science approach for improving multiple sequence alignment. PLoS One 7:e31362
77. Kwak D, Kam A, Becerra D et al (2013) Open-Phylo: a customizable crowd-computing platform for multiple sequence alignment. Genome Biol 14:R116
78. Alexandersson M, Cawley S, Pachter L (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res 13:496–502
79. Meyer IM, Durbin R (2002) Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics 18:1309–1318
80. Durbin R, Eddy S, Korgh A et al (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge
81. Sharma V, Elghafari A, Hiller M (2016) Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. Nucleic Acids Res 44(11):e103
82. Sinha S, He X (2007) MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLoS Comput Biol 3:e216
83. He X, Ling X, Sinha S (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. PLoS Comput Biol 5:e1000299
84. Majoros WH, Ohler U (2010) Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. PLoS Comput Biol 6: e1001037
85. Satija R, Pachter L, Hein J (2008) Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. Bioinformatics 24:1236–1242
86. Will S, Yu M, Berger B (2013) Structure-based whole-genome realignment reveals many novel noncoding RNAs. Genome Res 23:1018–1027
87. Earl D, Nguyen N, Hickey G et al (2014) Alignathon: a competitive assessment of whole-genome alignment methods. Genome Res 24:2077–2089
88. Darling ACE, Mau B, Blattner FR et al (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403
89. Edgar RC, Asimenos G, Batzoglou S et al (2011) Evolver: a whole-genome sequence evolution simulator. https://www.drive5.com/evolver
90. Dalquen DA, Anisimova M, Gonnet GH et al (2012) ALF–a simulation framework for genome evolution. Mol Biol Evol 29:1115–1123
91. Stoye J, Evers D, Meyer F (1998) Rose: generating sequence families. Bioinformatics 14:157–163
92. Cartwright RA (2005) DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics 21(Suppl 3):iii31–iii38
93. Pollard DA, Moses AM, Iyer VN et al (2006) Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. BMC Bioinform 7:376
94. Huang W, Nevins JR, Ohler U (2007) Phylogenetic simulation of promoter evolution: estimation and modeling of binding site turnover events and assessment of their impact on alignment tools. Genome Biol 8:R225
95. Varadarajan A, Bradley RK, Holmes IH (2008) Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. Genome Biol 9:R147

96. Fletcher W, Yang Z (2009) INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26:1879–1888
97. Kim J, Sinha S (2010) Towards realistic benchmarks for multiple alignments of non-coding sequences. BMC Bioinform 11:54
98. Arenas M, Posada D (2014) Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. Mol Biol Evol 31:1295–1301
99. Margulies EH, Cooper GM, Asimenos G et al (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res 17:760–774
100. Morgenstern B, Rinner O, Abdeddaïm S et al (2002) Exon discovery by genomic sequence alignment. Bioinformatics 18:777–787
101. Wang AX, Ruzzo WL, Tompa M (2007) How accurately is ncRNA aligned within whole-genome multiple alignments? BMC Bioinform 8:417
102. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A 87:2264–2268
103. Prakash A, Tompa M (2005) Statistics of local multiple alignments. Bioinformatics 21(Suppl 1):i344–i350
104. Prakash A, Tompa M (2007) Measuring the accuracy of genome-size multiple alignments. Genome Biol 8:R124
105. Frith MC, Park Y, Sheetlin SL et al (2008) The whole alignment and nothing but the alignment: the problem of spurious alignment flanks. Nucleic Acids Res 36:5863–5871
106. Landan G, Graur D (2007) Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol 24:1380–1383
107. Penn O, Privman E, Landan G et al (2010) An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27:1759–1767
108. Chang JM, Di Tommaso P, Notredame C (2014) TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol Biol Evol 31:1625–1637
109. Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100:659–674
110. Treangen TJ, Ondov BD, Koren S et al (2014) The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15:524
111. Frith MC, Hamada M, Horton P (2010) Parameters for accurate genome alignment. BMC Bioinform 11:80
112. Frith MC, Kawaguchi R (2015) Split-alignment of genomes finds orthologies more accurately. Genome Biol 16:106
113. Lunter G, Rocco A, Mimouni N et al (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. Genome Res 18:298–309
114. Herman JL, Novák Á, Lyngsø R et al (2015) Efficient representation of uncertainty in multiple sequence alignments using directed acyclic graphs. BMC Bioinform 16:108