Computation and Analysis of Genomic Multi-Sequence Alignments

Mathieu Blanchette

McGill Centre for Bioinformatics, McGill University, Montreal, Quebec, Canada, H3A 2B4; email: blanchem@mcb.mcgill.ca

Annu. Rev. Genomics Hum. Genet. 2007. 8:193–213

First published online as a Review in Advance on May 9, 2007.

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi:
10.1146/annurev.genom.8.080706.092300

1527-8204/07/0922-0193$20.00

Key Words

multiple sequence alignment, comparative genomics, whole-genome alignment, computational genome annotation, genome evolution

Abstract

Multi-sequence alignments of large genomic regions are at the core of many computational genome-annotation approaches aimed at identifying coding regions, RNA genes, regulatory regions, and other functional features. Such alignments also underlie many genome-evolution studies. Here we review recent computational advances in the area of multi-sequence alignment, focusing on methods suitable for aligning whole vertebrate genomes. We introduce the key algorithmic ideas in use today, and identify publicly available resources for computing, accessing, and visualizing genomic alignments. Finally, we describe the latest alignment-based approaches to identify and characterize various types of functional sequences. Key areas of research are identified and directions for future improvements are suggested.

Multiple sequence alignment (MSA): predicting homology of DNA bases from different genomes or from different regions of the same genome

Evolutionarily correct alignment: alignment where all and only homologous bases are aligned

INTRODUCTION

Although the first complete draft of the human genome was publicly released more than five years ago (75), its functional interpretation remains in its infancy. Much of the human DNA appears to be nonfunctional, and the role of most putatively functional regions remains poorly understood. One of the best ways to accurately annotate our genome (i.e., to delineate and characterize each type of functional element it contains) is through comparative genomics. By comparing our genome to that of other species and studying how different regions have evolved, one can often make educated guesses about the functions of these regions. This is one of the key motivations behind the sequencing of other vertebrate genomes. As of today, "complete" vertebrate genome sequences are available for human (75), chimpanzee (23), mouse (129), rat (57), dog (77), chicken (61), macaque (1), and Takifugu (3). Several other genome sequences are about to be released (including cow, African frog, opossum, zebrafish, Tetraodon) and others are in the process of being sequenced at a low coverage (2X) (see **http://www.genome.gov/10002154**). Large amounts of genomic sequence are also available for several species of insects (in particular fruit flies), worms, yeasts, plants, and a plethora of eubacteria and archebacteria.

To take advantage of this data for the purposes of doing genome annotation and evolutionary studies, these sequences must be meaningfully compared to each other, homologous regions identified, and their evolutionary history inferred. This is the problem of whole-genome sequence alignment, or, when more than two genomes are compared (which is when comparative genomics approaches gain most of their power), whole-genome multiple sequence alignment (MSA).

In this review, we give an overview of the latest algorithm advances for genomic multi-sequence alignment, discuss the limitations of these approaches, their accuracy evaluation, and their use for genome annotation purposes. We identify resources available for computing, accessing, and analyzing these alignments. Throughout, we outline key issues that remain to be addressed.

THE MULTIPLE SEQUENCE ALIGNMENT PROBLEM

MSA was one of the earliest topics studied by computational biologists (78, 109). Although the basic alignment principles apply to DNA, RNA, and amino acid sequences, specialized algorithms have been developed to address the particular challenges posed by each type of biomolecule. Here we focus on alignment of multiple, large (e.g., >1 Mb) genomic DNA sequences. Reviews of related topics can be found in Batzoglou (5), Dewey & Pachter (39), and Miller (88). For reviews of MSA of protein sequences, we refer the reader to Edgar & Batzoglou (46) and Wallace et al. (125).

Evolutionarily Correct Alignments

The input to the MSA problem usually consists of a set of genomic sequences, either completely assembled genomes, partially assembled draft genomes, or large finished genomic regions such as the ENCODE regions (49). Typically, the sequences being compared come from different species, but they can also be from the same species if they arose from duplication events. In many cases, a phylogenetic tree specifying the evolutionary relationships among the sequences being considered is adjoined to the input. This species tree might not accurately represent the history of every sequence region because of incomplete lineage sorting, gene duplications, gene conversion, or lateral gene transfer. If these confounding issues are temporarily ignored (as they mostly have been until recently), an evolutionarily correct MSA is an alignment where nucleotide i of species X is aligned to nucleotide j of species Y only if i and j are homologous, i.e., if, and only if, they were derived from a common ancestral nucleotide, either through faithful inheritance or through one or more substitutions. If the two nucleotides diverged because of a speciation event, they are called orthologous; they are called paralogous if their divergence results from a duplication [see Dewey & Pachter (39) for a more detailed classification]. Thus, one can see an MSA as a collection of sets of homologous nucleotides. This representation indirectly specifies the gaps in the alignment as positions where one sequence contains no nucleotide homologous to that of another sequence,[1] as a result of an insertion or a deletion (or incomplete data). An evolutionarily correct alignment thus provides much information about the evolutionary processes that led to the sequences considered. However, it does not unambiguously specify the history of all alignment bases, in particular because it does not differentiate between gaps due to insertions and those due to deletions (9, 24).

When duplications events are considered, defining the correct alignment becomes more difficult. The simple model of "one alignment row per species" fails to capture the possibility that several nucleotides from the same genome may be homologous. One solution is to allow each alignment column to contain more than one nucleotide per species, the goal being to identify any set of nucleotides (within species and across species) that share a common ancestry. Because no whole-genome MSA program published to date handles duplications in a completely satisfactory manner, we will postpone our discussion of this important aspect of evolution.

For several reasons, obtaining an evolutionarily correct MSA is not always possible. When sequences have diverged beyond a certain point, homologous bases cannot be reliably distinguished from nonhomologous ones. Information about homology is slowly but irremediably lost during evolution, and there is a limit to the accuracy of MSA algorithms. For example, very little noncoding DNA can be reliably aligned between human and fish because most of the vertebrate ancestral DNA that has not been deleted along one of the two lineages has usually diverged beyond recognition. Thus, the fact that a very small fraction of the human genome can be aligned to that of other nonmammalian species reflects a combination of the actual loss of homologous DNA and the increasing difficulty in reliably identifying it.

TFBS: transcription factor binding site

Function-Based Alignments

An alternative to searching for an evolutionarily correct alignment is to search for one that aligns regions that are functionally, but not necessarily evolutionarily, related. In species that are not too diverged (e.g., different mammals), functional homology often approximately equates to evolutionary homology, in particular for large, complex regions such as protein-coding genes. However, smaller functional regions such as transcription factor binding sites (TFBSs) can be functionally homologous without sharing a common ancestry. These functional regions are sufficiently short that they can arise from neutral drift and become fixed in a population because of selective pressure on the newly acquired function (37, 93). Many local multiple alignment programs (usually described as motif discovery algorithms) have been developed to identify such short functional regions. However, because the questions arising from this research area are quite distinct from those related to larger-scale evolutionary alignments, we refer the reader to Pavesi et al. (97) for more details. For the rest of this review, we focus on the identification, evaluation, and analysis of alignments in an evolutionary context.

Mathematical Objective Functions

Although the definition of an evolutionarily correct alignment is relatively simple, it

[1] It should be noted that this representation as sets of homologous nucleotides can often be depicted visually by several equivalent multiple sequence alignment (MSA) representations.

is not practical, as it is generally impossible to test. Instead, mathematical objective functions have been defined to approximate this notion. The most theoretically sound approach relies on stochastic models of sequence evolution specifying the probability of events such as substitutions, insertions, and deletions. Once such a model is defined, one evaluates a given MSA with its likelihood under the model by implicitly summing the probabilities of all evolutionary scenarios that are compatible with the alignment. The MSA sought is the one that maximizes this likelihood function. Although this idea is attractive from a theoretical standpoint, it quickly runs into practical difficulties. Evaluating the likelihood of a given MSA, even assuming a fixed tree topology and no duplications, is a NP-hard problem (24), i.e., one for which no efficient algorithm is believed to exist. For this reason, objective functions have historically been based on simpler schemes such as the sum-of-pairs score or a weighted version thereof (44). Today, most algorithms do not explicitly try to optimize any kind of global objective function, but simply follow a reasonable sequence of operations that usually produce relatively accurate alignments.

Global, Local, and “Glocal” Multiple Sequence Alignment

Alignment approaches have historically been divided into two groups: global and local. In a global alignment, the entirety of the sequences given as input are aligned in a single alignment, thus imposing the constraint that orthologous regions have to be colinear (i.e., no genome rearrangements and no duplications are allowed). If the sequences considered are closely related and are expected to have been completely derived from a common ancestor through a set of insertions, deletions, and substitutions, a global alignment is appropriate. In contrast, a local alignment predicts the homology of a set of fragments of the input sequences, leaving the rest of the sequences unaligned. In general, alignment programs identify a set of such local alignments, resulting in a mosaic of homology predictions, in which the segments aligned are not necessarily colinear in the input sequences. This enables one to find homologous segments that have undergone rearrangements. When a sizeable fraction of the sequences considered are expected to be highly diverged, a local alignment approach may also be preferable because by ignoring unalignable regions it reduces the risk of these regions causing misalignments of the whole sequence (26). However, the flexibility afforded by local alignment comes at a cost: By dropping ordering constraints, we increase the risk of aligning regions that are not true homologs (i.e., random hits). To maintain an acceptable level of false positives, one must accept a lower sensitivity.

The strategy adopted by most state-of-the-art alignment programs is a hybrid between local and global alignment [sometimes called a “glocal” alignment (16)] that takes advantage of the strengths of both methods. Most current MSA programs first identify sets of local alignments between pairs of sequences and then assemble the local alignments into chains of colinear sets of local alignments. Small local alignments that break the linearity of the surrounding regions are treated as likely false positives, thus improving specificity.

ALGORITHMS FOR MULTIPLE SEQUENCE ALIGNMENT

Algorithms for genomic MSA face a number of challenges, including:

1. **Computational complexity.** Optimal MSA is a NP-hard problem under most reasonable scoring schemes (47, 127), which means that the running time of all known exact algorithms for the problem is exponential in the number or the length of the sequences to be aligned.
2. **Large data sets.** Whole-genome sequence alignment programs have to be able to run on very large sequences. Despite the increasing computing power

available, this remains a very significant challenge and many compromises have to be made between accuracy and speed.

3. **Sensitivity.** Homologous nucleotides need to be identified for distantly related species.
4. **Specificity.** Nonhomologous nucleotides must not be aligned together.
5. **Model violations.** The evolutionary models and scoring schemes used by MSA programs are only approximations to reality. A good alignment program has to be robust with respect to these violations.

Different MSA programs make different trade-offs between these requirements, perhaps most obviously at the level of pair-wise alignment subroutines, upon which most MSA algorithms are based.

Seeded Pair-Wise Alignment

Most modern alignment methods rely on the ability to quickly and accurately align pairs of sequences or sequence "profiles" (see below). The accuracy of the multiple alignments produced depends in large part on that of the pair-wise alignment procedure used. Optimal local pair-wise alignments can be identified using the Smith-Waterman algorithm (117), but this approach is too slow for multimegabase sequences. For whole-genome alignments, nearly all existing approaches use a heuristic called seeded alignments, also at the base of the Blast algorithms (2), whereby local pair-wise alignments are explored only if they contain some type of highly conserved short match. Details vary from program to program: Blastz (110) explores exact matches over a set of nearby but nonconsecutive positions (called spaced seeds), LAGAN (15) relies on the CHAOS program (14) to identify short inexact (but consecutive) matches, AVID (12) searches for maximal exact matches (but allows mismatches at third codon positions), and MUMmer (34) uses suffix trees to identify maximal unique matches. The design of fast and sensitive seeded alignment algorithms remains a very active research area (79, 83, 121). When the aligned sequences are expected to largely consist of protein-coding regions (e.g., for prokaryote genomes), accuracy gains can be obtained by computing seeded pair-wise alignments at the translated level (35). For noncoding regulatory regions, Berezikov et al. (7) showed that sensitivity gains could be obtained by using seeds conserved for predicted TFBSs.

Seeded pair-wise alignment algorithm: heuristic used to align pairs of very large sequences, where local alignments are considered only if they contain a short, highly (and often perfectly) conserved match, called the seed

Once a good seed alignment is found, it is extended to the left and the right using various variants on the Smith-Waterman dynamic programming algorithm, this time allowing for gaps to be inserted. Note that at this stage the arrangement and orientation of pair-wise local alignments are unrestricted and one region can align to more than one other region, thus enabling one to find homology between sequences that have undergone genomic rearrangements and duplications.

Sets of pair-wise local alignments can be chained together into sets of colinear regions (63, 119, 134), using programs like Chaining/Netting (71), GRIMM-synteny (99), MAUVE (33), and MERCATOR (Colin Dewey, **http://www.biostat.wisc.edu/∼cdewey/mercator/**). Within a chain, more sensitive settings can be used to discover weaker pair-wise alignments located between anchors (31), and the process can be repeated in a recursive manner (15). However, such approaches are inapplicable for unassembled genomes [e.g., those sequenced at a 2X coverage (87)], thus posing significant new challenges to alignment algorithms.

Multiple Pair-Wise Alignments versus True Multiple Alignments

MSAs are often computed for the sole purpose of annotating one of the sequences (called the reference sequence—for example, the human sequence) using the others. The main interest is to find orthology relationships between regions of the reference sequence and those of the other species, but not to find the relationships between two nonreference species. In this case, why not simply compute pair-wise alignments between the human sequence and each other species and assemble these alignments into a human-centric multiple alignment? This was the strategy used by Multi-Pipmaker (111) and Vista (43), and it is adequate for closely related sequences. However, when sequences are more diverged, the MSA obtained in this way is less accurate than one obtained from an algorithm that attempts to identify orthology relationships among all species (21, 85). This is because correctly aligning two nonhuman sequences provides homology information that can help align them correctly to human. For this reason, most modern MSA programs instead use a phylogenetically guided progressive multiple alignment procedure.

Progressive Multiple Sequence Alignment

Many modern MSA programs are based on a simple scheme called "progressive alignment" (44), popularized by the Clustal package (22, 60). In a progressive alignment algorithm, a phylogenetic tree is first inferred on the basis of pair-wise alignments, and then used as a guide for the multiple alignment procedure (the tree can also be provided by the user). The two species that are the most closely related are aligned first, and the process is repeated until all sequences are aligned. Because the phylogenetic tree is binary, each step consists of the pair-wise alignment of one group of aligned sequences (corresponding to one clade) against another (corresponding to a sister clade), a process referred to as profile-profile alignment. Variants lie in the choice of algorithm and scoring function for profile-profile alignment. TBA (10) bases its profile-profile on pair-wise alignments of extant sequences. MLAGAN (15) uses a weighted sum-of-pairs approach, whereas MAVID (13) represents each aligned clade with its most likely common ancestral sequence. This last approach is the closest to the evolutionary view of multiple alignment. Whereas all these MSA algorithms use seed-based pair-wise sequence alignments at every step of the progressive alignment procedure, Flannick & Batzoglou (52) have proposed an algorithm that performs seed-based profile-profile alignment. By dynamically selecting more sensitive seeds in regions of each profile that are the most likely to yield matches (e.g., those corresponding to highly conserved regions), they report an improved sensitivity over standard seed-based approaches. Although the idea of progressive alignment was proposed more than 20 years ago, the devil is in the details, and the optimal trade-off between efficiency and accuracy has not yet been reached.

Other Multiple Alignment Strategies

Although progressive multiple alignment is intuitively appealing in an evolutionary context, nonphylogenetic alignment procedures have also been successfully developed. One of them, called consistency-based MSA, was pioneered by Morgenstern's group with the Dialign family of MSA programs (91, 101) and followed by others (122, 131). Like most progressive alignment approaches, the algorithm relies on the identification of pair-wise anchors (seed alignment). However, it weighs these pair-wise alignments based not only on the similarity of the two sequences involved, but also on the set of all other pair-wise anchors they are involved in with other species. Thus, if region R_x of species X has a weak alignment with region R_y of species Y, but both R_x and R_y have a strong alignment to region R_z of species Z, the reliability of the anchor alignment (R_x, R_y) is increased. Dialign follows this reweighting of anchors by a greedy procedure that attempts to identify the maximum weight set of collinear anchors. By using the entire set of anchors to decide which to retain in the final alignment, Dialign can afford a higher false-positive rate at the pair-wise level, and consequently a higher sensitivity. Recently, Do et al. (40) integrated a probabilistic version of the consistency-based approach into a progressive MSA algorithm for protein sequences called ProbCons. Paten & Birney (96) extended this promising idea to genomic MSA in the program PECAN.

In a different direction, Zhang & Waterman (133) formulated the local MSA problem as one identifying Eulerian paths in a graph, a problem for which efficient algorithms exist, and used this approach to identify new families of functional elements and retrotransposons. Raphael et al. (105) improved the approach to make it suitable for the alignment of highly diverged, rearranged, repetitive genomic regions, and the results are competitive with more traditional progressive alignment methods.

Refining Alignments

To compute alignments reasonably quickly for large genomes, genomic MSA programs have to cut corners, which sometimes results in suboptimal alignments. For example, with a progressive MSA approach, the subalignment produced for a certain clade is not revised later in the light of other sequences, which sometimes leads to nonoptimal alignments. Although the overall structure of the alignment produced may often be correct, details may not be. Many authors have proposed postprocessing algorithms to refine multiple alignments—a process also known as realignment (20). Most approaches assume that the initial alignment is approximately correct and attempt to fine-tune the position of gaps. By assuming global correctness, computational complexity is reduced so more sensitive approaches can be used. A simple procedure proposed by several authors and implemented in CLUSTALW and MLAGAN is to repeatedly remove individual sequences from the MSA and realign them to the remaining n-1 aligned sequences. When the data set considered contains closely related sequences, a more effective approach might be to remove a complete subtree (e.g., the primate clade) and realign it to the remaining sequences. Other approaches include making local changes to the alignment by moving gaps to optimize some objective function (for example, using a genetic algorithm) (126).

Uncertainty in Alignments

One drawback of the fast alignment algorithms described above is that they do not provide an assessment of the uncertainty related to particular regions of the alignment. Some global alignment methods (e.g., MLAGAN) attempt to align every nucleotide of every sequence, and others leave some regions unaligned, but none quantitate confidence in the alignment correctness. A small step in that direction is a procedure called "alignment cleaning" (81), which identifies alignment regions that do not meet certain statistical significance thresholds (104), and then either dissolves these blocks or breaks them into highly supported subalignments.

Meaningful confidence measures of alignment accuracy require a probabilistic model of sequence evolution. One can then ask questions like: Given a set of sequences $X_1, \ldots, X_n$, what is the probability that nucleotide i of sequence X_a is orthologous to nucleotide j of sequence X_b? This type of question is addressed by a new type of alignment problem called statistical alignment. Although the existing algorithms for statistical MSA (59) remain too slow to be executed on a large scale, they bear the promise of a richer representation of MSA, with important applications in phylogeny, functional sequence annotation, and ancestral genome reconstruction.

BENCHMARKING MULTIPLE SEQUENCE ALIGNMENT PROGRAMS

Objective and quantitative benchmarking procedures are necessary to evaluate existing MSA algorithms, point out weaknesses, and suggest improvements. MSA benchmarking is challenging because the evolutionary correctness of a MSA cannot be tested for actual biological sequences. Furthermore, the mathematical scoring schemes described above are only approximations to the biological quality of an alignment, so using them as the basis of MSA evaluation might be misleading. For protein MSA, additional sources of information such as 3D structure and functional assays can be used to help determine the correct alignment; databases such as BaliBase (124) and Prefab (45) contain large numbers of such alignments against which protein MSA programs can be evaluated. Because of the lack of external data indicating the correct alignment, no such database exists for genomic sequences.

Two types of benchmarking approaches have been proposed for genomic MSA. In the first, a set of well-annotated orthologous genomic sequences are considered, and the quality of the MSA is assessed based on the fraction of the annotated functional regions that are correctly aligned. For example, Bray & Pachter (13) have used known coding exons as landmarks. A drawback of this approach is that well-annotated functional regions (e.g., coding exons) are often the easiest to align properly because of their relatively large size, high degree of conservation, and preserved linearity of the gene structure. Thus, estimates based on the success of aligning such regions may be optimistic. Evaluations based on other types of functional regions (e.g., regulatory regions) have also been considered (55) but remain problematic because of the lack of well-annotated noncoding functional regions. However, large-scale, unbiased experimental data from projects such as ENCODE (49) are now providing an excellent basis for this type of benchmarking (86).

The second type of benchmarking involves synthetic sequences (102). Here the evolution of a synthetic DNA sequence is computationally simulated along the branches of a given phylogenetic tree. By keeping track of the mutations performed along each branch, the evolutionarily correct alignment can be constructed. The leaf sequences (synthetic extant sequences) are then fed to the alignment program being evaluated, and the alignment obtained is compared to the correct one. Various approaches have been proposed to compare the two alignments (10, 89), but most revolve around the evaluation of the sensitivity (fraction of orthologous sites that are correctly aligned) and specificity (fraction of aligned sites that are orthologous) of the predicted alignment. The usefulness of such benchmarking depends on the realism of the sequence evolution simulation and of the synthetic sequence chosen for the root. For example, Blanchette et al. (10) developed a fairly realistic simulation of neutrally evolving sequences, including variable mutation rates along different lineages as well as retrotransposon insertions, and used it to compare their MSA program with others. A shortcoming of their approach is that actual DNA sequences contain a mixture of slow- and fast-evolving sites, something that is not modeled in their simulation. More sophisticated evolution simulators allowing both neutrally evolving and constrained elements, as well as tandem and segmental duplications and DNA polymerase slippage, are thus necessary for an improved assessment of alignment accuracy.

Whether benchmarking is done on real or synthetic sequences, the evaluation of alignment accuracy has to be carried out keeping in mind the context in which the alignment will be used. If the alignment is computed to seek out highly conserved functional regions such as coding regions one may care less about the accuracy of the alignment in neutrally evolving regions and more about the sensitivity of local alignments. However, if the alignment is used to identify weaker signals such as TFBSs, the accuracy of the alignment on noncoding regions is key (32). Finally, for genome evolution studies and ancestral genome reconstruction efforts (9), the alignment accuracy for neutrally evolving regions is paramount, and modeling the effects of alignment errors on the analysis is important (82, 106).

USING MULTIPLE SEQUENCE ALIGNMENT FOR GENOME ANNOTATION

One of the most powerful approaches to separate functional from nonfunctional genomic regions and identify the putative role of the former is through the analysis of the "evolutionary signatures" of functional regions. Nonfunctional regions mostly evolve neutrally and tend to accumulate mutations at rates that are relatively well understood (27). Functional regions are under selective pressure to preserve their function.[2] Here we briefly review the types of functional sequence analysis approaches enabled by MSA. This is only a partial list and virtually every aspect of computational genome annotation has benefited from the availability of multiple alignments [see (42) for a related treatment of the topic].

Identification of Regions Under Selective Pressure

A large fraction of many metazoan genomes is believed to be nonfunctional in the sense that it does not encode a beneficial function for the host other than perhaps maintaining an appropriate spacing between functional elements. For example, by comparing the human and mouse genomes, Waterston et al. (129) estimated that only about 5% of the DNA of each genome is under selection. One of the original incentives to sequence more mammalian genomes was to help identify and characterize these functional regions.

Many approaches have been developed to scan a genomic MSA and identify regions under selection. Most of these methods are phylogenetically aware, i.e., they use an evolutionary model and a phylogenetic tree to evaluate sequence conservation. Some use a sliding window approach (54) and perform a hypothesis test to try to reject the null model of neutral evolution [e.g., binCons (84)]. Others estimate position-specific substitution rates and identify regions under selection using a hidden Markov model (HMM) approach [PhastCons (115), eShadow (94)] or some other flexible window approach [GERP (30)]. These have the advantage of being able to detect long, weakly conserved regions as well as short, highly conserved ones. Various genome browsers offer precomputed conservation scores for whole-genome alignments (e.g., PhastCons on the UCSC Genome Browser and GERP on the Ensembl Genome Browser). Although the methodological differences between the various approaches are important, the agreement among their predictions tends to be rather high, and the main source of uncertainty in the identification of conserved regions remains the difference between alignment programs (86). A number of genomic regions have been dissected and analyzed based on MSA (e.g., 6, 21, 65, 86), generally resulting in the identification of many more constrained regions than can be accounted for by protein-coding selective pressure. Although the precise identification of regions under selective pressure, especially shorter ones, remains a challenge, we believe that the more promising research direction now lies in the interpretation of evolutionary signatures to predict the putative function of each region.

Evolutionary signature: pattern of interspecies sequence conservation/variation typical of a particular type of function (protein-coding gene, RNA gene, regulatory region, etc.)

Identifying Protein-Coding Regions

Although the constitutive coding exons of most human genes have now been accurately identified, much uncertainty remains regarding the number and location of alternatively spliced exons, especially those rarely expressed and thus poorly represented in transcript databases. Alignment-based gene prediction has proven to be a useful complement to gene annotation based on experimental data because the evolutionary signature of coding exons is clear, strong, and well understood. High-synonymous to nonsynonymous

[2] This, however, does not mean that their sequence is perfectly or even highly conserved as a whole, but rather that its functional features tend to be.

substitution rate ratios, absence of in-frame stop codons, and conserved splice sites are signatures sought by a number of gene-finding algorithms, using either pair-wise (90, 132) or MSA (18, 38, 58). These algorithms generally far outperform single-organism gene finders. Because coding exons tend to be well conserved throughout vertebrates but most other types of functional and nonfunctional regions are not, vertebrate multiple alignment is extremely informative for human gene finding, although species- or clade-specific exons will systematically be missed by such approaches. All of these approaches depend on an accurate MSA, and Dewey et al. (38) have proposed to circumvent this problem by performing joint alignment (using a translated seed-based alignment) and gene prediction.

Identifying RNA Genes

Structural RNA genes (tRNAs, rRNAs, microRNA precursors, etc.) also have a specific evolutionary signature that can be detected based on an accurate MSA. A number of RNA gene predictors attempt to detect compensatory mutations in RNA molecules (those compensating one substitution by another, to preserve Watson-Crick complementarity), the most recent being Evofold (98, 128) and RNAz (41, 128). However, RNA gene prediction presents very specific challenges as the primary sequence often lacks detectable conservation, resulting in the difficulty of obtaining accurate MSA. It was suggested very early on (108) that a better way to proceed might be to jointly predict multiple alignment and RNA secondary structure. However, the algorithms proposed thus far (114) remain too slow to be used on a large scale.

Identifying Regulatory Regions

One of the most promising uses of MSA is in identifying transcriptional regulatory regions, an approach coined "phylogenetic footprinting." Generally located in noncoding regions, TFBSs often stand out by their interspecies conservation. Thus, the simplest approach is to use sequence conservation as a preliminary filter before searching for TFBSs using position weight matrix scans [e.g., ConSite (107)]. Most binding sites are flexible and do not require perfect sequence conservation to remain functional, and many recent approaches based on position weight matrix scanning [rVISTA (80), PreMod (51), and others] use conservation of estimated binding affinity rather than raw sequence conservation, or TFBS-specific evolutionary models [Monkey (92)] to detect putative aligned binding sites. However, it is now recognized that binding site turnover, in which random mutations create and destroy binding sites, is quite common (37), and that lack of conservation does imply lack of function (32). There is clearly a need for more accurate models of binding-site turnover and for algorithms using them for binding-site prediction [see (93) for a good step in this direction].

De novo motif discovery algorithms have also started taking advantage of genomic multiple alignments, either to favor the identification of motifs whose occurrences tend to be conserved across species (70, 103, 113, 116), or to use reliable alignment anchors as constraints for the identification of shorter conserved motifs (11, 50). Finally, we note that mammalian regulatory regions appear to have an evolutionary signature that can be detected even without explicitly referring to TFBSs, as was demonstrated by the Hardison group (48, 72, 73, 123).

Genome Evolution Studies

Because they attempt to recapitulate the evolutionary relationships between genomes, genomic multi-sequence alignments are at the core of many phylogenomics studies (36, 70, 82, 100, 118) and genome evolution studies (71, 130). These alignments also underly the interesting prospect of accurately reconstructing the sequence of ancestral genomes from the genome of extant species (9, 24, 64).

Although phylogenetic studies are outside the scope of this review, we warn the reader against the pitfalls of phylogenetic inference and evolutionary studies based on dubious alignments. When dealing with sets of sequences whose correct alignment is not obvious, alignments are often sensitive to particular parameters values of the alignment program (e.g., gap open and extend penalties), and downstream analyses may be biased (27, 28, 106). One elegant solution to this problem is statistical alignment, which allows sampling alignments from their posterior probability distribution to estimate evolutionary distance and trees (62, 82).

RESOURCES

Despite the high quality of alignment programs available today, constructing large multiple alignments remains a challenging project that requires expertise and substantial computing power. Although the tools available are becoming increasingly accurate, they are generally not trivial to use and often involve the execution of a number of intermediate steps that are not always well described. For this reason, most whole-genome MSA users rely on publicly available precomputed alignments or on alignment Web servers for analysis. Here we briefly review those that are most commonly used (see also 53).

Precomputed Genomic Multi-Sequence Alignments

Because of the computational power and infrastructure required, whole-genome multiple alignment remains out of reach for the typical biology laboratory. However, for the most studied genomes, whole-genome multiple alignments have been computed using one or more of the programs discussed above and are being made available to the community. **Table 1** lists a set of whole-genome MSAs currently available for various groups of species. In many cases, alignments are integrated into genome browsers such as the UCSC Genome Browser (74) and the Ensembl Genome Browser (8), which enable visualization of the alignment itself, as well as of many other types of sequence annotations.

Although computation of large MSAs requires an impressive feat of computational engineering, handling and analyzing them are also not simple tasks. Whole-genome alignments files often consist of several gigabytes

Table 1 Publicly available precomputed whole-genome alignments

Species	Resource	MSA tool(s) used	URL
- 17 vertebrate genomes - ENCODE regions - 9 insect genomes - 2 nematode genomes	UCSC Genome Browser	TBA/MULTIZ	**http://genome.ucsc.edu/**
- 9 vertebrate genomes	Ensembl Genome Browser	PECAN	**http://www.ensembl.org/index.html**
- 5 vertebrate genomes - 3 plant genomes - 2 Ciona genomes - 2 Phytophtora genomes	Vista Browser	Shuffle-LAGAN	**http://pipeline.lbl.gov/cgi-bin/gateway2**
- 3 nematode genomes	Wormbase	MLAGAN	**http://wormbase.org/**
- 9 vertebrate genomes - 6 *Drosophila* genomes	ECR Browser	Blastz	**http://ecrbrowser.dcode.org/**
- 4 yeast genomes	Broad Institute	Kellis et al. (70)	**http://www.broad.mit.edu/annotation/**
- 12 *Drosophila* genomes	Eisen Lab	Mercator/MAVID, MAUVE, MLAGAN, MULTIZ	**http://rana.lbl.gov/drosophila/**

Table 2 Genomic multiple sequence alignment (MSA) Web servers

MSA tool	URL
CLUSTALW (22, 60)	**http://www.ebi.ac.uk/clustalw**
MAVID (13)	**http://baboon.math.berkeley.edu/mavid**
CHAOS/DIALIGN (14)	**http://dialign.gobics.de**
MAFFT (69)	**http://timpani.genome.ad.jp/∼mafft/server**
MLAGAN	**http://lagan.stanford.edu/lagan_web/index.shtml**
Mulan (95)	**http://mulan.dcode.org**
aliWABA (66)	**http://aba.nbcr.net**

of data, and extracting meaningful information from them requires efficient tools. Several groups provide computer source code or Web servers to handle and analyze large alignments. Web-based tools include the UCSC Table Browser (67) and Galaxy (56), whereas popular open-source software includes Jim Kent's source tree (**http://hgwdev.cse.ucsc.edu/∼kent**), a set of C programs at the core of the UCSC Genome Browser, and the EnsMart/BioJava/BioPerl (68) code underlying some of the Ensembl Genome Browser. The latter has the advantage of being able to remotely query Ensembl data sets without having to install them on a local machine, although this approach runs into efficiency problems for complex or numerous queries.

Multiple Sequence Alignment Web Servers

All multiple alignment programs mentioned in this review can be freely downloaded (for academic use) and can be run locally. However, in many cases (see **Table 2**), there are Web servers that compute alignments for a set of user-provided sequences, thus avoiding the hassle of software installation while taking advantage of dedicated machines optimized for that purpose. Note, however, that input sequences are usually limited to a few megabases. The SinicView program (112) allows the simultaneous computation and comparison of MSAs computed using several of the tools listed in **Table 2**.

Alignment Visualization

Visualization of multiple alignments is crucial for developing an intuitive understanding of the evolution of a particular genomic region, for integrating information from different sources, and for detecting possible alignment errors. A number of tools have been developed to address different aspects of visualization. Genome browsers such as the UCSC Genome Browser (74) and the Ensembl Genome Browser (8) are designed to visually integrate information from different sources, including multiple alignment. These browsers are based on the visualization of data related to a given reference sequence and typically exclude or indirectly represent bases that do not align to bases in the reference sequence.

One drawback of these tools is that it is currently not possible to visualize user-provided alignments or alignments computed on the fly from user-provided sequences. Tools allowing visualization and manipulation of user-provided MSA are listed in **Table 3** (programs analyzing only pair-wise alignments are omitted). Most visualization tools perform basic sequence analysis tasks such as the detection of conserved regions, and most allow the display of user-provided annotation. Many allow variable zoom-in resolution.

FUTURE CHALLENGES

The Holy Grail of the genomic MSA community is to design efficient algorithms to

Table 3 Visualization tools for multiple sequence alignments (MASs)

Visualization Tool	Functionality	URL
ABC (29)	Multiresolution views; user-provided annotations; conserved region detection	**http://mendel.stanford.edu/sidowlab/**
VISTA family (17, 43)	Visualization of alignment of user-provided sequence against precomputed MSA; multiresolution views; user-provided annotation; conserved region detection	**http://genome.lbl.gov/vista/index.shtml**
K-Browser (19)	True MSA view (i.e., not projected on a reference sequence); visualization of genome annotation on multiple genomes simultaneously	**http://hanuman.math.berkeley.edu/kbrowser**
COMPAM (76)	Visualization of multiple pair-wise alignments; automatic COG-based gene annotation; conserved regions detection and clustering; synteny analysis	**http://bio.informatics.indiana.edu/projects/compam/**
MaM (1)	Automated functional annotation of MSA; motif detection; extraction of subalignments based on annotation	**http://compbio.cs.sfu.ca/MAM.htm**
JalView (25)	MSA editing; conserved region detection; basic phylogenetic inference. Mostly designed for protein sequences	**http://www.jalview.org**

identify maximum likelihood MSA based on a realistic multiscale probabilistic model of sequence evolution that allows for substitutions, insertions, deletions, retrotransposon activity, duplications, and genome rearrangements. Although various groups have been developing increasingly accurate sequence evolution models for each type of mutation, these models have not yet been integrated into genomic MSA tools. For example, there appears to be much to gain from using a context-dependent insertion and deletion model of DNA polymerase slippage (4). However, at some point, remote homology detection becomes impossible with a generic model of evolution and requires multiple alignment algorithms tailored to specific types of functional regions [e.g., SLAM (38) for coding regions, or CONREAL (7) for regulatory regions]. We expect that improvement in alignment accuracy will come from a combination of fast, existing approaches and slower, domain-specific methods (e.g., 120).

We believe that one of the keys to progress in the MSA area is the development of algorithms that explicitly quantify the uncertainty in the alignment produced. This would allow the identification of more speculative alignments, which could then be included or excluded at will by the user. However, representing uncertainty is not simply a matter of providing a number describing the probability of correctness of a given region of the alignment, but also requires producing and representing the many possible alternate solutions. Genome annotation tools based on MSA should be developed that take this uncertainty into consideration.

Finally, we believe that MSA-based genome annotation approaches (in particular, RNA gene finding, discovery of regulatory regions, etc.) are only starting to take full advantage of the evolutionary signatures contained in these alignments. The large quantity of high-quality data coming out of projects like the ENCODE project (49) will allow the development of improved evolutionary models on which the next generation of computational genome annotation algorithms will be based.

SUMMARY POINTS

1. Genomic MSAs aim to recapitulate evolutionary history by identifying the regions that share a common ancestry in genomes of different species.
2. Because of the computational complexity of the problem and the very large size of the genomes to be aligned, computer programs have to trade off accuracy for efficiency. However, for most recent algorithms, the loss in accuracy is minimal.
3. Accurate genomic multiple alignments greatly facilitate the computational prediction of protein-coding genes, RNA genes, and regulatory regions. This is achieved by detecting patterns of sequence conservation that are characteristic of a particular type of function.
4. Many of the best MSA programs are run as Web servers and their results can be analyzed using interactive visualization tools. Pre-computed whole-genome multiple alignments can be downloaded for most groups of sequenced genomes.

FUTURE ISSUES TO BE RESOLVED

1. Genomic multi-sequence alignment remains an open problem. Even the best programs available are fairly slow, have a relatively poor sensitivity for diverged sequences, and often erroneously align nonhomologous nucleotides. Algorithms using more realistic evolutionary models are likely to improve the situation.
2. Benchmarking alignment of programs remains challenging. Functional and comparative genomics projects such as ENCODE are now providing data to evaluate and improve alignment programs.
3. Each type of functional region evolves under a different type of selective pressure that induces a different type of evolutionary signature. Characterizing and detecting such signatures for noncoding functional regions are keys to improving computational genome annotation.
4. Alignment tools representing the uncertainty in the alignment being reported need to be developed, and this uncertainty needs to be considered by alignment-based computational annotation approaches.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Alkan C, Tuzun E, Buard J, Lethiec F, Eichler EE, et al. 2005. Manipulating multiple sequence alignments via MaM and WebMaM. *Nucleic Acids Res.* 33:W295–98
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–10

3. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297:1301–10
4. Ball EV, Stenson PD, Abeysinghe SS, Krawczak M, Cooper DN, Chuzhanova NA. 2005. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.* 26:205–13
5. Batzoglou S. 2005. The many faces of sequence alignment. *Brief Bioinform.* 6:6–22
6. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–25
7. Berezikov E, Guryev V, Plasterk RH, Cuppen E. 2004. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res.* 14:170–78
8. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. 2006. Ensembl 2006. *Nucleic Acids Res.* 34:D556–61
9. Blanchette M, Green ED, Miller W, Haussler D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* 14:2412–23
10. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14:708–15
11. Blanchette M, Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* 12:739–48
12. Bray N, Dubchak I, Pachter L. 2003. AVID: a global alignment program. *Genome Res.* 13:97–102
13. Bray N, Pachter L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* 14:693–99
14. Brudno M, Chapman M, Gottgens B, Batzoglou S, Morgenstern B. 2003. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinform.* 4:66
15. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13:721–31
16. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, et al. 2003. Glocal alignment: finding rearrangements during alignment. *Bioinformatics* 19(Suppl. 1):i54–62
17. Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, et al. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.* 14:685–92
18. Carter D, Durbin R. 2006. Vertebrate gene finding from multiple-species alignments using a two-level strategy. *Genome Biol.* 7(Suppl. 1):S6 1–12
19. Chakrabarti K, Pachter L. 2004. Visualization of multiple genome annotations and alignments with the K-BROWSER. *Genome Res.* 14:716–20
20. Chakrabarti S, Lanczycki CJ, Panchenko AR, Przytycka TM, Thiessen PA, Bryant SH. 2006. State of the art: refinement of multiple sequence alignments. *BMC Bioinform.* 7:499
21. Chapman MA, Donaldson IJ, Gilbert J, Grafham D, Rogers J, et al. 2004. Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. *Genome Res.* 14:313–18
22. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31:3497–500
23. Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87

24. Chindelevitch L, Li Z, Blais E, Blanchette M. 2006. On the inference of parsimonious indel evolutionary scenarios. *J. Bioinform. Comput. Biol.* 4:721–44

25. Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. *Bioinformatics* 20:426–27

26. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301:71–76

27. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* 13:813–20

28. Cooper GM, Brudno M, Stone EA, Dubchak I, Batzoglou S, Sidow A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* 14:539–48

29. Cooper GM, Singaravelu SA, Sidow A. 2004. ABC: software for interactive browsing of genomic multiple sequence alignment data. *BMC Bioinform.* 5:192

30. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15:901–13

31. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, et al. 2003. Strategies and tools for whole-genome alignments. *Genome Res.* 13:73–80

32. Deleted in proof

33. Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14:1394–403

34. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. 1999. Alignment of whole genomes. *Nucleic Acids Res.* 27:2369–76

35. Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* 30:2478–83

36. Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6:361–75

37. Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19:1114–21

38. Dewey C, Wu JQ, Cawley S, Alexandersson M, Gibbs R, Pachter L. 2004. Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res.* 14:661–64

39. Dewey CN, Pachter L. 2006. Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Hum. Mol. Genet.* 15(Spec. No. 1):R51–56

40. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.* 15:330–40

41. Dowell RD, Eddy SR. 2006. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinform.* 7:400

42. Down TA, Hubbard TJ. 2004. What can we learn from noncoding regions of similarity between genomes? *BMC Bioinform.* 5:131

43. Dubchak I, Ryaboy DV. 2006. VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods Mol. Biol.* 338:69–89

44. Durbin R, Eddy S, Krogh A, Mitchison G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge Univ. Press

45. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* 5:113

46. Edgar RC, Batzoglou S. 2006. Multiple sequence alignment. *Curr. Opin. Struct. Biol.* 16:368–73
47. Elias I. 2006. Settling the intractability of multiple alignment. *J. Comput. Biol.* 13:1323–39
48. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* 13:64–72
49. ENCODE Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306:636–40
50. Fang F, Blanchette M. 2006. FootPrinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.* 34:W617–20
51. Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F, Blanchette M. 2006. PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.* 35(Database issue): D122–26
52. Flannick J, Batzoglou S. 2005. Using multiple alignments to improve seeded local alignment algorithms. *Nucleic Acids Res.* 33:4563–77
53. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC. 2003. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.* 13:1–12
54. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32:W273–79
55. Giardine B, Elnitski L, Riemer C, Makalowska I, Schwartz S, et al. 2003. GALA, a database for genomic sequence alignments and annotations. *Genome Res.* 13:732–41
56. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15:1451–55
57. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521
58. Gross SS, Brent MR. 2006. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* 13:379–93
59. Hein J, Wiuf C, Knudsen B, Moller MB, Wibling G. 2000. Statistical alignment: computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.* 302:265–79
60. Higgins DG, Sharp PM. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73:237–44
61. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716
62. Holmes I. 2005. Using evolutionary Expectation Maximization to estimate indel rates. *Bioinformatics* 21:2294–300
63. Huang X, Chao KM. 2003. A generalized global alignment algorithm. *Bioinformatics* 19:228–33
64. Hudek AK, Brown DG. 2005. Ancestral sequence alignment under optimal conditions. *BMC Bioinform.* 6:273
65. Hughes JR, Cheng JF, Ventress N, Prabhakar S, Clark K, et al. 2005. Annotation of *cis*-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc. Natl. Acad. Sci. USA* 102:9830–35
66. Jones NC, Zhi D, Raphael BJ. 2006. AliWABA: alignment on the web through an A-Bruijn approach. *Nucleic Acids Res.* 34:W613–16

67. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32:D493–96
68. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, et al. 2004. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* 14:160–69
69. Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33:511–18
70. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–54
71. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100:11484–89
72. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. 2005. Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res.* 15:1051–60
73. Kolbe D, Taylor J, Elnitski L, Eswara P, Li J, et al. 2004. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.* 14:700–7
74. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, et al. 2006. The UCSC genome browser database: update 2007. *Nucleic Acids Res.* 35(Database issue):D668–73
75. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
76. Lee D, Choi JH, Dalkilic MM, Kim S. 2006. COMPAM: visualization of combining pairwise alignments for multiple genomes. *Bioinformatics* 22:242–44
77. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–19
78. Lipman DJ, Altschul SF, Kececioglu JD. 1989. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA* 86:4412–15
79. Lippert RA, Zhao X, Florea L, Mobarry C, Istrail S. 2005. Finding anchors for genomic sequence comparison. *J. Comput. Biol.* 12:762–76
80. Loots GG, Ovcharenko I. 2004. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.* 32:W217–21
81. Loytynoja A, Milinkovitch MC. 2001. SOAP, cleaning multiple alignments from unstable blocks. *Bioinformatics* 17:573–74
82. Lunter G, Miklos I, Drummond A, Jensen JL, Hein J. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinform.* 6:83
83. Ma B, Tromp J, Li M. 2002. PatternHunter: faster and more sensitive homology search. *Bioinformatics* 18:440–45
84. Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* 13:2507–18
85. Margulies EH, Chen CW, Green ED. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet.* 22:187–93
86. Margulies EH, Cooper G, Asimenos G, Thomas DJ, Dewey CN, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* In press
87. Margulies EH, Vinson JP, Miller W, Jaffe DB, Lindblad-Toh K, et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci. USA* 102:4795–800

88. Miller W. 2001. Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* 17:391–97

89. Morgenstern B, Goel S, Sczyrba A, Dress A. 2003. AltAVisT: comparing alternative multiple sequence alignments. *Bioinformatics* 19:425–26

90. Morgenstern B, Rinner O, Abdeddaim S, Haase D, Mayer KF, et al. 2002. Exon discovery by genomic sequence alignment. *Bioinformatics* 18:777–87

91. Morgenstern B, Werner N, Prohaska SJ, Steinkamp R, Schneider I, et al. 2005. Multiple sequence alignment with user-defined constraints at GOBICS. *Bioinformatics* 21:1271–73

92. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* 5:R98

93. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. 2006. Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLOS Comput. Biol.* 2:e130

94. Ovcharenko I, Boffelli D, Loots GG. 2004. eShadow: a tool for comparing closely related sequences. *Genome Res.* 14:1191–98

95. Ovcharenko I, Loots GG, Giardine BM, Hou M, Ma J, et al. 2005. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.* 15:184–94

96. Paten B, Birney E. 2006. Pecan. **http://www.ebi.ac.uk/∼bjp/pecan/**

97. Pavesi G, Mauri G, Pesole G. 2004. In silico representation and discovery of transcription factor binding sites. *Brief Bioinform.* 5:217–36

98. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLOS Comput. Biol.* 2:e33

99. Pevzner P, Tesler G. 2003. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13:37–45

100. Philippe H, Blanchette M. 2007. Overview of the First International Conference on Phylogenomics. *BMC Evol. Biol.* **7**(Suppl 1):S1

101. Pohler D, Werner N, Steinkamp R, Morgenstern B. 2005. Multiple alignment of genomic sequences using CHAOS, DIALIGN and ABC. *Nucleic Acids Res.* 33:W532–34

102. Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB. 2004. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinform.* 5:6

102a. Pollard DA, Moses AM, Iyer VN, Eisen MB. 2006. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinform.* 7:1–18

103. Prakash A, Blanchette M, Sinha S, Tompa M. 2004. Motif discovery in heterogeneous sequence data. *Pac. Symp. Biocomput.* 348–59

104. Prakash A, Tompa M. 2005. Statistics of local multiple alignments. *Bioinformatics* 21(Suppl. 1):i344–50

105. Raphael B, Zhi D, Tang H, Pevzner P. 2004. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Res.* 14:2336–46

106. Rosenberg MS. 2005. Multiple sequence alignment accuracy and evolutionary distance estimation. *BMC Bioinform.* 6:278

107. Sandelin A, Wasserman WW, Lenhard B. 2004. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.* 32:W249–52

108. Sankoff D. 1985. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* 45:810–25

109. Sankoff D, Cedergren RJ. 1983. Simultaneous comparison of three or more sequences related by a tree. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. Sankoff D and Kruskal JB, pp. 253–63. Reading, MA: Addison-Wesley

110. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* 13:103–7

111. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, et al. 2000. PipMaker–a web server for aligning two genomic DNA sequences. *Genome Res.* 10:577–86

112. Shih AC, Lee DT, Lin L, Peng CL, Chen SH, et al. 2006. SinicView: a visualization environment for comparisons of multiple nucleotide sequence alignment tools. *BMC Bioinform.* 7:103

113. Siddharthan R, Siggia ED, van Nimwegen E. 2005. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLOS Comput. Biol.* 1:e67

114. Siebert S, Backofen R. 2005. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 21:3352–59

115. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034–50

116. Sinha S, Blanchette M, Tompa M. 2004. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinform.* 5:170

117. Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–97

118. Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.* 59:191–209

119. Sobel E, Martinez HM. 1986. A multiple sequence alignment program. *Nucleic Acids Res.* 14:363–74

120. Stocsits RR, Hofacker IL, Fried C, Stadler PF. 2005. Multiple sequence alignments of partially coding nucleic acid sequences. *BMC Bioinform.* 6:160

121. Sun Y, Buhler J. 2006. Choosing the best heuristic for seeded alignment of DNA sequences. *BMC Bioinform.* 7:133

122. Szklarczyk R, Heringa J. 2006. AuberGene–a sensitive genome alignment tool. *Bioinformatics* 22:1431–36

123. Taylor JTS, King DC, Hardison RC, Miller W, Chiaromonte F. 2006. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res.* 16:1596–604

124. Thompson JD, Koehl P, Ripp R, Poch O. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* 61:127–36

125. Wallace IM, Blackshields G, Higgins DG. 2005. Multiple sequence alignments. *Curr. Opin. Struct. Biol.* 15:261–66

126. Wang C, Lefkowitz EJ. 2005. Genomic multiple sequence alignments: refinement using a genetic algorithm. *BMC Bioinform.* 6:200

127. Wang L, Jiang T. 1994. On the complexity of multiple sequence alignment. *J. Comput. Biol.* 1:337–48

128. Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* 102:2454–59

129. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62

130. Yang S, Smit AF, Schwartz S, Chiaromonte F, Roskin KM, et al. 2004. Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res.* 14:517–27

131. Ye L, Huang X. 2005. MAP2: multiple alignment of syntenic genomic sequences. *Nucleic Acids Res.* 33:162–70

132. Zhang L, Pavlovic V, Cantor CR, Kasif S. 2003. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res.* 13:1190–202

133. Zhang Y, Waterman MS. 2005. An Eulerian path approach to local multiple alignment for DNA sequences. *Proc. Natl. Acad. Sci. USA* 102:1285–90

134. Zhang Z, Raghavachari B, Hardison RC, Miller W. 1994. Chaining multiple-alignment blocks. *J. Comput. Biol.* 1:217–26

Annual Review of
Genomics and
Human Genetics

Volume 8, 2007

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters may be found at http://genom.annualreviews.org/