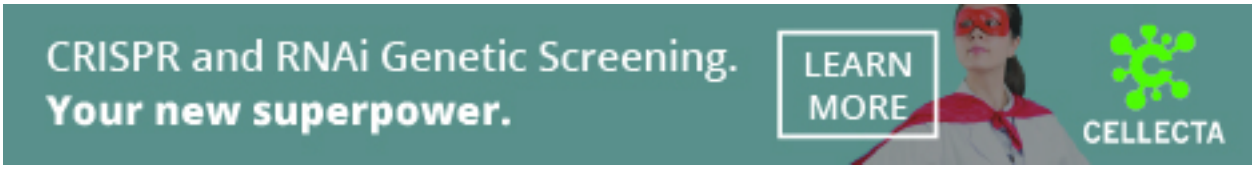
Uncertainty in homology inferences: Assessing and improving genomic sequence alignment

Gerton Lunter, Andrea Rocco, Naila Mimouni, et al.

Genome Res. 2008 18: 298-309 originally published online December 11, 2007
Access the most recent version at doi:10.1101/gr.6725608



Methods

Uncertainty in homology inferences: Assessing and improving genomic sequence alignment

Gerton Lunter,[1,3] Andrea Rocco,[2] Naila Mimouni,[2] Andreas Heger,[1] Alexandre Caldeira,[2] and Jotun Hein[2]

[1]*MRC Functional Genetics Unit, University of Oxford, Department of Physiology, Anatomy, and Genetics, Oxford OX1 3QX, United Kingdom;* [2]*Department of Statistics, University of Oxford, Oxford Centre for Gene Function, Oxford, OX1 2TG, United Kingdom*

Sequence alignment underpins all of comparative genomics, yet it remains an incompletely solved problem. In particular, the statistical uncertainty within inferred alignments is often disregarded, while parametric or phylogenetic inferences are considered meaningless without confidence estimates. Here, we report on a theoretical and simulation study of pairwise alignments of genomic DNA at human–mouse divergence. We find that >15% of aligned bases are incorrect in existing whole-genome alignments, and we identify three types of alignment error, each leading to systematic biases in all algorithms considered. Careful modeling of the evolutionary process improves alignment quality; however, these improvements are modest compared with the remaining alignment errors, even with exact knowledge of the evolutionary model, emphasizing the need for statistical approaches to account for uncertainty. We develop a new algorithm, Marginalized Posterior Decoding (MPD), which explicitly accounts for uncertainties, is less biased and more accurate than other algorithms we consider, and reduces the proportion of misaligned bases by a third compared with the best existing algorithm. To our knowledge, this is the first nonheuristic algorithm for DNA sequence alignment to show robust improvements over the classic Needleman–Wunsch algorithm. Despite this, considerable uncertainty remains even in the improved alignments. We conclude that a probabilistic treatment is essential, both to improve alignment quality and to quantify the remaining uncertainty. This is becoming increasingly relevant with the growing appreciation of the importance of noncoding DNA, whose study relies heavily on alignments. Alignment errors are inevitable, and should be considered when drawing conclusions from alignments. Software and alignments to assist researchers in doing this are provided at http://genserv.anat.ox.ac.uk/grape/.

[Supplemental material is available online at www.genome.org.]

Most, if not all, of comparative genomics relies crucially on the quality of sequence alignments. As a consequence, the sequence-alignment problem has received a great deal of attention. However, despite having been introduced over three decades ago (Needleman and Wunsch 1970), it remains an active area of research (for reviews, see Batzoglou 2005; Dewey and Pachter 2006). One reason for this continued interest is that alignments produced by existing algorithms still show considerable disagreement (Dewey et al. 2006). This disagreement is often thought to result from differences in the algorithm's accuracy due to, e.g., inaccurate evolutionary models or suboptimal choices of parameters (Waterman et al. 1992; Gusfield et al. 1994; Elofsson 2002; Dewey et al. 2006). Here, we argue instead that alignment accuracy is more fundamentally limited. Rather than resulting from inaccurate models or parameters, differences between inferred alignments may simply reflect uncertainties resulting from the limited information available in extant sequences, from which different algorithms infer distinct but equally plausible homologies (Lassmann and Sonnhammer 2005). To the extent that this is true, attention should be directed toward quantifying this unavoidable uncertainty rather than toward optimizing the evolutionary model underpinning the algorithm. Quantifying this uncertainty will help experimentalists assess which alignment regions can be relied upon in subsequent analyses.

Several authors have considered uncertainty in alignments. Byers and Waterman looked at the problem of enumerating suboptimal alignments (Waterman 1983; Byers and Waterman 1984), but this approach proved impractical because of the sheer number of such alignments. An alternative approach focuses instead on reliable individual columns within alignments. A "conditional best score" can be computed for alignments that include any particular residue pair (Sellers 1979; Goad and Kanehisa 1982; Altschul and Erickson 1986; Zuker 1991). Given an arbitrary threshold, these scores delineate regions of homology in a dot plot rather than a single best alignment. A similar approach has been used to calculate a "reliability index" for individual pairings (Chao et al. 1993), which has been used to improve alignments (Mevissen and Vingron 1996; Schlosshauer and Ohlsson 2002; Tress et al. 2003). Despite such improvements, alignment quality remains a major issue, e.g., for homology modeling of protein structure (Tramontano et al. 2001). One difficulty is that proteins evolve under selection, which is hard to model, so that any simulation must necessarily be highly idealized. Lacking realistic simulated data, alignment algorithms must be calibrated using databases of structural alignments (Mevissen and Vingron 1996; Tress et al. 2003; Edgar and Batzoglou 2006), which are limited in size and accuracy and biased toward globular proteins.

[3]**Corresponding author.**
E-mail gerton.lunter@dpag.ox.ac.uk; fax 44-1865-282651.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6725608.

18:298–309 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

A second issue is that, with notable exceptions (Do et al. 2005; Paten and Birney 2007), alignment algorithms are almost universally based on optimizing a score rather than on a probabilistic model. Aside from parameterization issues, this makes it difficult to interpret the score and reliability indices derived from it, which has hampered the rational design of novel algorithms based on these statistics.

Here, we focus on the probabilistic alignment of mammalian genomic DNA. The importance of this problem is underlined by the increasing number of available genomes and the requirement of full-length alignments, particularly for the comparative study of conserved noncoding DNA. These elements are embedded in large amounts of neutrally evolving sequence, which, in many cases, retain sufficient sequence identity to be alignable (Waterston et al. 2002). This allows a different strategy from that used for protein alignments; rather than modeling evolution under functional constraint, neutral evolution may be modeled to optimize the alignment of the neutral majority of sequence (Chiaromonte et al. 2002). The conserved fraction, being easier to align, may be processed by the same evolutionary model. Because of the large amount of available data, the process of neutral evolution is known in great detail (Waterston et al. 2002; Hwang and Green 2004; Meunier and Duret 2004; Hellmann et al. 2005; Lunter et al. 2006), allowing the simulation of realistic sequence pairs whose homologies are known exactly, an approach that was profitably used to assess the quality of fly genome alignments (Pollard et al. 2004).

We emphasize that in this study we consider only part of the whole-genome alignment problem: the pairwise local alignment of homologous nucleotide sequences. We ignore the issues of finding anchors and of dealing with repetitive sequence, genomic inversions and duplications, and nonorthologous relationships (Blanchette et al. 2004; Bray and Pachter 2004; Brudno et al. 2004; Dewey and Pachter 2006; Sun and Buhler 2006), which are crucial, but can be separated from the nucleotide-level alignment problem. Here, we thus assume that regions of homology are reliably assigned (see Prakash and Tompa 2007 for a recent study considering this issue), and we focus on the problem of inferring nucleotide-level homology. We also do not consider the multiple-alignment problem, which is essentially more difficult than pairwise alignments. However, a detailed understanding of the issues concerning pairwise alignments will, we hope, help guide the design of probabilistic multiple-alignment algorithms.

A central observation made in this study is that alignment errors follow particular patterns and cause alignments to be biased in particular ways. Depending on the application, it is important to be aware of (and account for) the type and extent of these biases. For instance, naïve estimates of indel rates are systematically negatively biased, and explicit accounting for alignment biases greatly reduces their impact (Lunter 2007). We distinguish three types of alignment error, termed gap wander (Holmes and Durbin 1998), gap attraction, and gap annihilation. We show that, to varying degrees, all probabilistic and score-based aligners tested exhibit these biases. For the most prevalent of these, gap wander, we obtain an analytic expression of its contribution to alignment error.

Having established that alignment errors are prevalent, we turn to probabilistic alignment algorithms. A key advantage of probabilistic aligners is their ability to assign posterior probabilities to individual alignment columns (Thorne et al. 1991; Durbin et al. 1998; Metzler 2003; Lunter et al. 2005). We show that this posterior probability accurately predicts the true probability that individual columns are correct. This suggests that, rather than using a standard maximum-likelihood approach such as the Viterbi algorithm, posteriors could be profitably used to identify good alignments. Posterior decoding-alignment algorithms were proposed some time ago (Krogh 1997; Durbin et al. 1998; Holmes and Durbin 1998), and recently there has been a renewed interest in probabilistic alignment algorithms, mostly focusing on proteins (Do et al. 2005; Kall et al. 2005; Roshan and Livesay 2006; Paten and Birney 2007), and similar approaches were found to improve RNA folding (Ding et al. 2005). In contrast, the performance of posterior decoding algorithms on genomic DNA sequences has, to our knowledge, not been investigated in detail before. Here, we examine two novel posterior decoding algorithms and find that they show superior performance compared with the standard Viterbi decoding and with score-based aligners, in terms of sensitivity and the extent of alignment biases.

Although our models and algorithms improve upon earlier alignment algorithms, a key point of this study is to emphasize that errors in alignments are unavoidable. We show that this is true even when the underlying evolutionary model and parameters are known exactly. For sequences whose divergence is comparable to human and mouse, we recover 83%–88% of homologous residue pairs, depending on the model and the decoding algorithm, stressing the need to quantify the remaining uncertainty in the alignment. The best-performing model explicitly accounts for variations in GC content, and the particular form of the mammalian indel-length spectrum; surprisingly, modeling the variation in substitution and indel rates themselves had little, if any, effect on the resulting alignment. Independently of the model used, the posterior-decoding algorithms were found to be superior to the Viterbi algorithm. By comparison, of the score-based aligners used in the ENCODE project, BLASTZ (Schwartz et al. 2003; Blanchette et al. 2004) shows the best overall performance, achieving a sensitivity of 82%, similar to the sensitivity of Viterbi alignments.

Despite the advantages of a probabilistic approach, most aligners currently are score based rather than probabilistic. One reason is that probabilistic algorithms are perceived to be more complex. It is therefore important to emphasize that the algorithms we used have the same asymptotic time and memory complexity as standard score-based algorithms; for example, the Viterbi algorithm (Durbin et al. 1998) is formally identical to the Needleman–Wunsch algorithm (Needleman and Wunsch 1970). Based on the ideas presented here, we have developed a probabilistic genome aligner, GRAPe, which we used to compute human and mouse genome alignment. The software will be described more fully elsewhere and is available at http://genserv.anat.ox.ac.uk/grape. A genome browser for the human–mouse alignments is accessible through the same URL.

Results

Biases in alignment

Alignment algorithms, whether probabilistic or score based, compute alignments that are systematically biased. We describe the three most important biases, in order of their frequency of occurrence, and discuss their effects on alignments.

The most frequent cause of misaligned bases is due to an effect termed “gap wander” or “edge wander” (Holmes and Durbin 1998). Gap wander occurs because the mutation process creates random and spurious local sequence similarities, which

compete with the sequence similarities due to homology (Fig. 1A). Long unrelated sequences are unlikely to show similarities comparable to those of homologous sequences, but short regions of similarity do frequently occur and cannot be distinguished from true homology. As a result, the most likely location of a gap often differs from its true location. Gap wander causes alignment columns near gaps to show an inflated average sequence similarity while simultaneously causing the proportion of columns that are correctly aligned (the alignment accuracy) to be lowest near to gaps.

Under a Jukes–Cantor model of evolution, it is possible to investigate the effects of gap wander analytically. Following, in part, the argument by Holmes (1998) and Holmes and Durbin (1998), we find the proportion of misaligned bases due to gap wander to be

$$F_w = \frac{\sigma}{\gamma}\frac{(4e^{4\sigma/3}-3)(4e^{4\sigma/3}-1)}{8e^{4\sigma/3}-3}, \qquad (1)$$

valid for small divergences (see Appendix A). Here, γ is the ratio of the substitution rate, σ, to the indel rate, δ. Note that the dominant term in (1) is linear in σ, in fact $F_w = (3\sigma/5\gamma) + O(\sigma^2)$, and for this reason, we say that gap wander is a first-order effect.

The second most prevalent bias is termed "gap attraction." This is an interaction effect between indels, and occurs when two indels hit homologous sequences at nearby positions. In this case, the most parsimonious explanation often involves one rather than two gaps, even at the cost of additional substitutions. Because this additional cost is, in expectation, proportional to the distance between the gaps, the result is an apparent "attraction" between gaps (see Fig. 1B,C). It causes a downward bias in the number of inferred indels, and further decreases the alignment accuracy near gaps. Since gap attraction is an interaction effect, the number of affected sites in alignments is of second order in the divergence.

The third bias, "gap annihilation", is also an interaction effect between indels, but occurs at lower frequencies. When two indels have identical length but are of opposing signature (e.g., an insertion followed by a nearby deletion in the same lineage; or two deletions in separate lineages), the evolutionary history competing with the true explanation involves no indels altogether (Fig. 1D). Since indels are relatively rare, this explanation is favored even when it requires a considerable number of additional inferred substitutions. The evolutionary scenarios causing this situation may sound contrived, but in fact, the probability that two indels have identical length is ~20% in human–mouse alignments, because most gaps are short (~30% are single-nucleotide indels). The results of gap annihilation are, again, a downward bias in the number of inferred indels, a reduction of the alignment accuracy, and a decrease in the apparent sequence similarity. These effects occur nearly uniformly across the alignment (see Lunter 2007 for a more detailed discussion), in contrast to the biases induced by gap wander and gap annihilation, which strongly colocalize with inferred gaps.

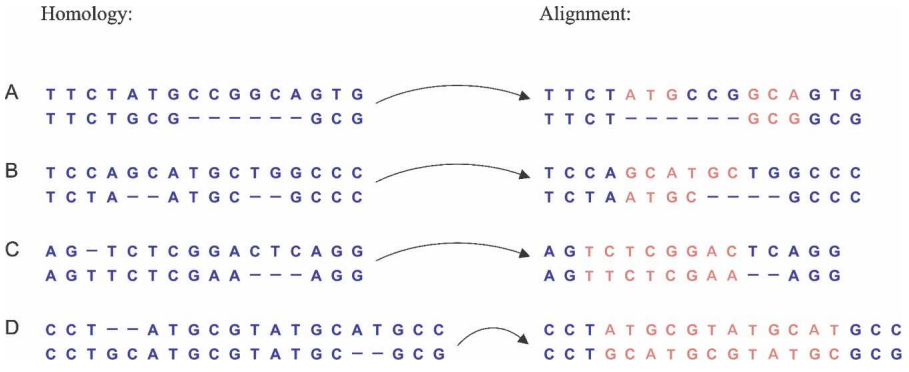


Figure 1. Three types of alignment bias. Alignment algorithms are consistently biased toward likely distributions of indels across sequences, despite the occurrence of less-likely configurations at low frequencies. The figure shows four pairs of sequences with their homologies (*left*) and corresponding most-likely alignments (*right*), with wrongly aligned bases highlighted. We distinguish between three types of bias: gap wander (*A*), caused by spurious high-sequence similarity at nonhomologous sites; gap attraction (*B,C*), occurring when two indels have little separation, and gap annihilation (*D*), which occurs when two indels of equal size but opposite signature are found near to each other, favoring explanations without indel events.

Simulation study of alignment biases

To show that the three types of biases influence alignments as predicted, we designed a simulation study. We generated sequences so that "true" homologies were known, after which we removed gaps and realigned the resulting sequences. We evolved sequences under the Jukes–Cantor model (Jukes and Cantor 1969) with $\sigma = 0.375$ expected substitutions per site (corresponding to an average sequence identity of 0.705), and we used a geometric indel model with substitution/indel rate ratio of $\gamma = 7.5$ (see Methods section). These parameters result in sequences that are comparable to sequence at human–mouse divergence (mean sequence identity 69%). The simulated sequences were realigned under the same model using Viterbi decoding. The use of the Jukes–Cantor model allowed comparisons with our analytical result (1). Simulations show that sequences evolved and aligned under the HKY model (Hasegawa et al. 1985) show very similar alignment biases (see Supplemental Table S1 and Supplemental Fig. S1).

We find that the alignment accuracy is lowest for columns adjacent to gaps, as predicted, with only 56% aligned correctly (see Fig. 2A). The apparent average sequence identity for these columns is 85%, much higher than the true sequence identity, 70.5%. Both observations are compatible with the combined action of gap wander and gap attraction.

Moving away from gaps, the apparent sequence identity quickly drops to nearly the correct value and continues to decrease to ~68%. In contrast, the accuracy rises slowly and plateaus at around 96% far away from gaps. This again agrees with our predictions, since all alignment biases act to decrease the accuracy at medium distances from gaps, while gap attraction and gap wander have opposite effects on sequence identity. In balance, gap wander dominates near gaps, while at medium distances, their effects nearly cancel. The fact that neither sequence identity nor alignment accuracy reach optimal values in gap-distal regions reflects the effects of gap annihilation, which is the dominant alignment bias away from alignment gaps. Gap attraction, finally, is responsible for the scarcity of closely spaced gaps (Fig. 2B).

We next investigated the dependence of alignment accuracy with sequence divergence. The analytic predic-

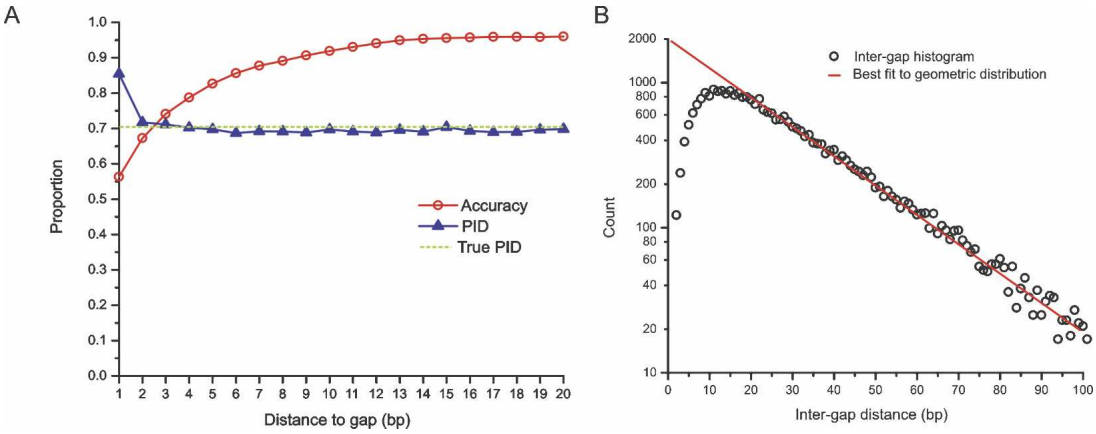


Figure 2. Effects of alignment biases in relation to gaps. Alignment biases cause systematic errors in alignments that are non-uniformly distributed with respect to alignment gaps. (*A*, *left*) The proportion sequence identity (PID, blue triangles), the true PID (dashed), and the proportion of correctly aligned columns (accuracy, red circles), for realigned sequences evolving under a Jukes–Cantor model, as a function of the distance to the nearest gap in the inferred alignment. The spuriously high PID and low accuracy adjacent to gaps is caused by gap wander. Gap annihilation is responsible for the reduced accuracy, and the slight reduction of PID below the true value away from gaps. (*B*, *right*) A histogram of intergap distances (circles), and the best fit to a geometric distribution (red line). The scarcity of closely spaced gaps (less than about 20 nucleotides apart) is due to gap attraction and affects a large number of gaps (note the logarithmic scale).

tion (1) of the false-positive fraction (FPF, see Methods section for a definition) closely agrees with the observations ($\sigma = 0.075$, FPF = $F_w = 0.008$; $\sigma = 0.150$, FPF = $F_w = 0.022$; $\sigma = 0.225$, FPF = 0.047, and $F_w = 0.041$). Since gap wander is the only bias considered in the analysis, the very good agreement indicates that gap wander is the dominant cause of alignment error in the low-divergence regime. For higher divergences, the observed FPF exceeds the predicted value (Fig. 3) because second-order indel interactions such as gap annihilation become more prevalent, as indicated by the reduction in asymptotic accuracy.

To investigate the effect of inaccurate parameters on alignment quality, we realigned all simulated sequences with fixed parameters rather than with those used in the simulation. This has little negative effect on alignment quality (Supplemental Fig. S2). Given the strong biases present in alignments, could it be that detuning the parameters might actually improve alignments? For instance, decreasing the gap penalty would increase the number of gaps, opposing the bias in gap density due to gap attraction. To investigate this, we again simulated sequences under a Jukes–Cantor model, and realigned them using Viterbi decoding with a model parameterized by a range of substitution and indel rates (Fig. 4). Sensitivity is maximal (84%) when parameters coincide with the simulation parameters, both for the indel and the substitution rates. However, the sensitivity stays within 1% of the maximum across a wide range of indel rates ($0.02 \leq \delta \leq 0.10$) and substitution rates ($0.20 \leq \sigma \leq 0.45$). We conclude that alignment quality is robust against fairly large errors in the values of the evolutionary parameters.

Model fidelity and alignment accuracy

The initial simulations show that for parameters corresponding to human–mouse alignments, only 84% of homologous bases were aligned correctly. Because of the simple model, this result may not be representative for actual human–mouse alignments. To make a more realistic assessment of the expected quality of such alignments, we developed a test set of simulated sequences that accurately approximates evolution along the human and mouse lineages. To assess the impact of model fidelity, we realigned this data set using a hierarchy of models and inferred alignments using three different decoding algorithms for each model in turn.

We simulated evolution using parameters that closely mimic human–mouse evolution. Specifically, we simulated the following aspects: large-scale variation of GC content; GC-content-dependent indel rates; an empirical indel-length spectrum; dependence of the substitution model on GC content; and GC-independent local substitution rate variation. The evolutionary parameters were obtained from BLASTZ human–mouse alignments (see Methods section for details). In all, we simulated 20,000 sequence pairs with an average length of 700 nt, with

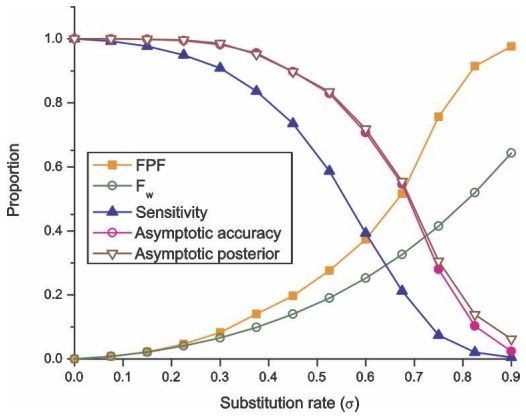


Figure 3. Dependence of alignment accuracy on evolutionary distance. Accuracy decreases with increasing evolutionary distance. Shown are the false-positive fraction (FPF, orange squares); the predicted FPF based on gap wander alone (F_w, green open circles); the sensitivity (blue solid triangles); the proportion of correct alignment columns at distance 15 from the nearest gap (asymptotic accuracy, pink dots); and the average posterior probability at the same distance (asymptotic posterior, brown open triangles). Sequences were simulated for various values of the divergence σ (horizontal axes), and realigned using the same σ value. The substitution/indel rate ratio was fixed at $\gamma = \sigma/\delta = 7.5$. Qualitatively the same behavior is seen when realigning using a fixed σ (see Supplemental Fig. S2).

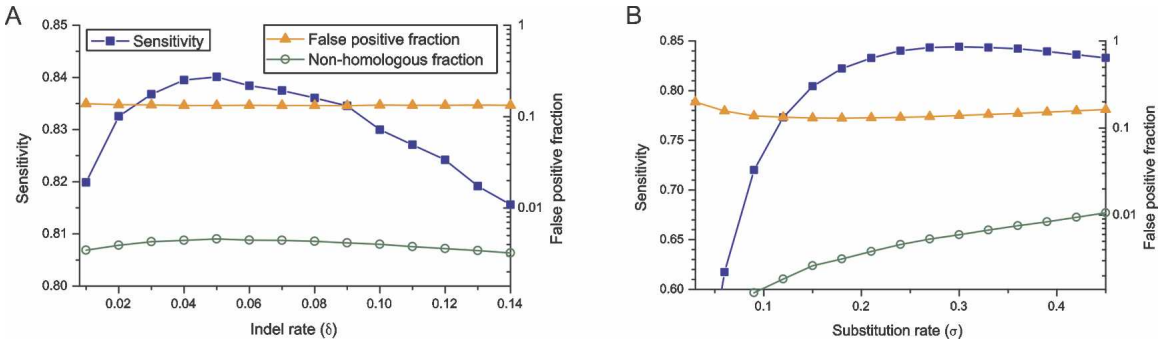


Figure 4. Suboptimal parameters have minimal impact on alignment accuracy. Shown are: sensitivity to identify homologous nucleotide pairs (blue squares, on *left* axis), the false-positive fraction (orange triangles, on *right* axis), and the nonhomologous fraction (green circles, on *right* axis). Sequences were generated under a Jukes–Cantor model with substitution rate $\sigma = 0.3$ *and indel rate* $\delta = 0.05$, and (*A*) realigned using a fixed substitution rate $\sigma = 0.3$ and a range of indel rates, and (*B*) using a fixed indel rate $\delta = 0.05$ and variable σ.

2×100 nt of flanking sequence added as appropriate for local alignments. Note that this presents a realistic scenario for a whole-genome aligner when a fairly dense set of anchors has been generated.

The simulated sequences were then realigned using a hierarchy of probabilistic aligners (Table 1; Fig. 5A). The most elaborate ("Full") model tracked all of the evolutionary-rate variation used in the simulations. In addition, this model uses a geometric mixture model to closely approximate the empirical indel-length spectrum (Fig. 5B). The other models were obtained by allowing only one parameter to vary, while other parameters were fixed to their average values (see Table 1). Finally, we considered a "Basic" model, obtained by pegging all parameters to their averages and replacing the indel-length model by a standard geometric distribution, corresponding to affine gap penalties.

For each model, we compared three decoding algorithms to infer alignments from the sequence data. As baseline method, we used the standard Viterbi decoding algorithm, which computes the single most likely alignment that is compatible with the observed sequences. In addition, we used two posterior decoding algorithms, referred to here as posterior decoding and marginalized posterior decoding (MPD; see Appendix B for details). Both algorithms compute the alignment that maximizes the cumulative log posterior probability of all contributing alignment columns. This is equivalent to maximizing the product of column posteriors (Fariselli et al. 2005) and has the advantage of removing the need for arbitrary gap weighting to account for variable lengths of alignments, which is required for standard sum-of-posteriors decoding (Durbin et al. 1998; Do et al. 2005; Kall et al. 2005; Roshan and Livesay 2006).

We summarized the results using three summary statistics: sensitivity, false-positive fraction (FPF), and nonhomologous fraction (NHF; see Methods section for definitions) (Fig. 6). Of the inference procedures, MPD achieves the best sensitivity (88.1%), with good FPF (13.3%) and NHF (1.6%). Viterbi alignments are more conservative, resulting in a notably lower sensitivity (84.9%), but slightly better performance on the FPF and NHF statistics (12.7% and 0.38%). Standard posterior decoding shows comparable sensitivity (87.9%), but high FPF and NHF scores (14.3% and 2.3%).

Beside good sensitivity and FPF ratings, MPD also shows fewer alignment biases. The average PID next to gaps is only mildly elevated at 72.9%, compared with 80.1% for Viterbi, indicating a reduced impact of gap wander. Gap attraction is also less prevalent, as indicated by the distribution of distances between gaps, which is closer to the ideal geometric distribution (Fig. 7). Finally, MPD alignments show a high asymptotic accuracy (97.3% accurate at distance 15 from gaps, compared with 96.1% for Viterbi alignments), suggesting a reduced impact of gap annihilation. This reduced impact of alignment biases improves the estimate of the number of indels (0.0394 gaps per nucleotide for MPD, compared with 0.0345 for Viterbi), although a substantial bias remains (true gap density, 0.0490 gaps/nt). All statistics mentioned are for the Full model; there appears to be little interaction between the model and the inference procedure, and the conclusions remain valid across the model hierarchy.

Table 1. A hierarchy of probabilistic alignment models

Model	δ[a]	α[b]	ε_1[c]	ε_2[c]	Substitution probabilities
Basic	0.0486	1.0	0.652	-	avg.[d]
VarIndel	fGC-dep[e]	1.0	0.652	-	avg.
VarSubs	0.0486	1.0	0.652	-	local divergence dependent[f]
MixtureGeometric	0.0486	0.857	0.652	0.906	avg.
SequenceContent	0.0486	1.0	0.652	-	fGC-dependent
Full	fGC-dep	fGC-dep	fGC-dep	fGC-dep	dependent on fGC and local divergence

[a]Indel rate parameter.
[b]Mixture coefficient of mixture geometric indel-length distribution.
[c]Parameters of geometric indel-length distribution.
[d]Signifies that average parameters are used (Supplemental Table S2).
[e]Signifies that parameters depend on the fraction GC (see Supplemental Table S2).
[f]Signifies that substitution rate parameters are tuned to the local sequence divergence (see Supplemental Table S3). All explicit parameter values are averages (see Supplemental Table S2).

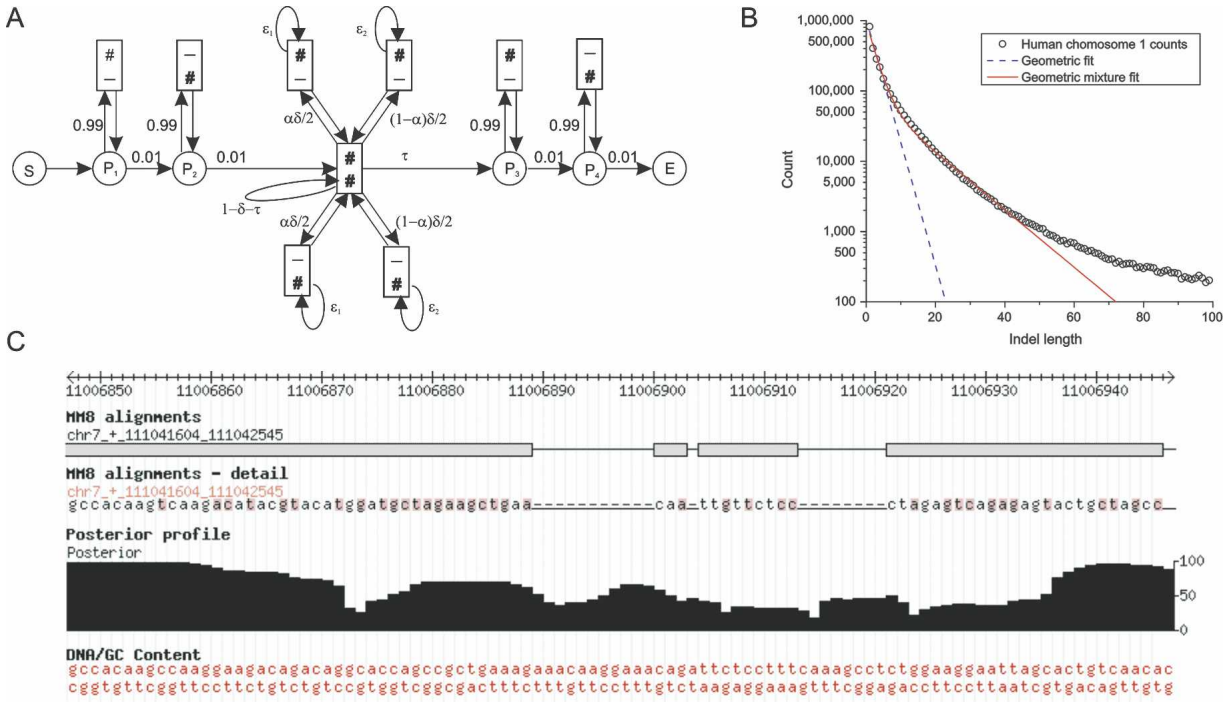


Figure 5. Topology of the pair HMM for probabilistic alignments. (*A*) The model is implemented as a pair HMM with a match state (center) surrounded by delete (*top*) and insert (*bottom*) states. Hash signs (#) signify emissions, dashes (–) represent no emission (rather than the emission of a gap character); circles represent silent states and are included for clarity, and arrows represent allowed transitions. Paths through this HMM correspond to alignments (and dash signs then represent gap characters). Local alignments were computed by surrounding the core HMM by two pairs of "padding" states (P_1 to P_4) allowing the alignable portion of the sequences to be embedded in nonhomologous sequence. Note that the model allows a single pass through the central pair HMM, and padding sequence is allowed at both ends of the alignment only. (*B*) The observed indel-length spectrum in BLASTZ human–mouse alignments (*right*, circles) is better approximated by a mixture of two geometric distributions (red solid line) than by a single geometric distribution (corresponding to affine-gap scores; blue dashed line). This mixture distribution is implemented by duplicating the insert and delete states. Parameters of the model are: δ, the indel probability per aligned site; ε_1 and ε_2, the parameters governing the indel length distribution; α, the geometric mixture coefficient, τ, the alignment length parameter. (*C*) Screenshot of the alignment browser, showing a marginalized posterior decoding (MPD) alignment computed using this model, together with posterior column probabilities. Alignments generally contain columns with low posterior probability, indicating regions where competing alignments contribute a significant fraction of the total likelihood.

We were surprised to find that increasing the model complexity has little effect on the performance. For the MPD alignments, the FPF varies between 13.34% and 15.34%, the NHF varies in the range of from 1.45% to 3.19%, and the sensitivity ranges from 87.56% to 88.22%. Compared with the Basic model, and irrespective of the decoding algorithm used, the models that vary either local substitution rates (VarSubs) or indel rates (VarIndel) show little or no improvement in any of the three statistics. This is consistent with our finding that Jukes–Cantor alignments are robust to variations in evolutionary rate parameters. Tuning the substitution model to the sequence GC content improves the FPF (13.6%, from 15.2%) and the NHF (1.45%, from 2.85%), but also somewhat reduces the sensitivity (87.6%, from 87.8%). Modeling the indel-length spectrum using a geometric mixture model has the opposite effect of increasing the sensitivity (to 88.2%) at the cost of an increased NHF (3.19%), while the FPF improves (15.1%), but only slightly, compared with the Basic model (15.2%). The Full model strikes a good balance with the best FPF (13.3%) and good NHF and sensitivity scores (1.63% and 88.1%; Fig. 7).

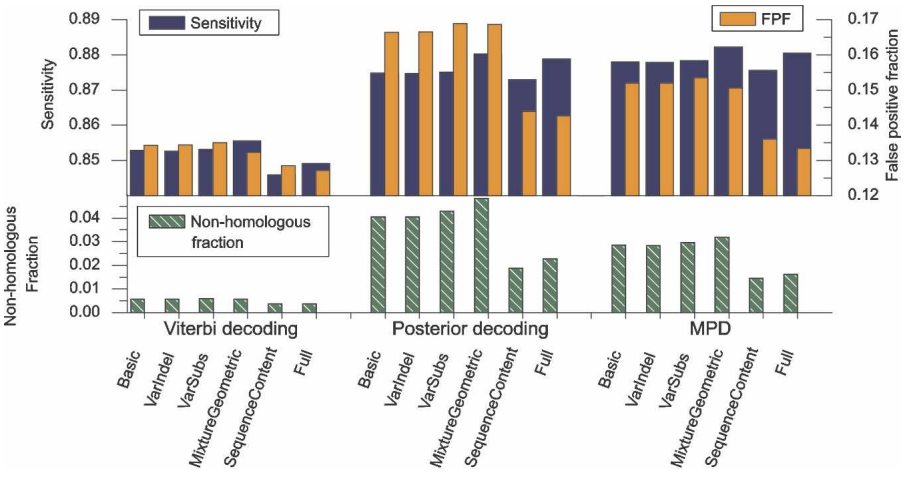


Figure 6. Dependence of alignment accuracy on modeling fidelity and inference procedure. Shown are the sensitivity, false-positive fraction, and nonhomologous fraction for three inference algorithms and various alignment models (see Table 1) used to align sequences from the human–mouse evolutionary simulation.

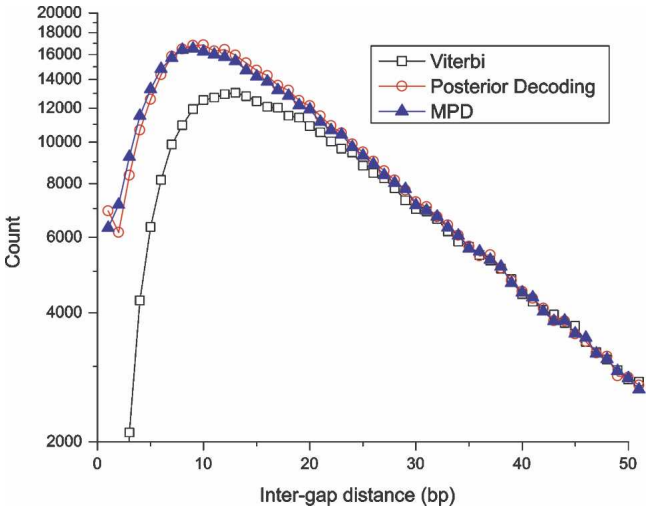


Figure 7. Posterior decoding shows fewer alignment biases. Shown are the intergap distance histograms for alignments obtained by Viterbi decoding (open squares), posterior decoding (open circles) and MPD (filled triangles), applied on the Full model. The scarcity of closely spaced gaps, resulting from gap attraction, is apparent for all decoding algorithms, but is much less pronounced for posterior decoding and MPD than for Viterbi decoding.

Our test setup implicitly assumes that alignment algorithms can use the correct evolutionary parameters, which is not true in practice. For this reason, our results should be regarded as providing an upper limit to the achievable alignment accuracy for the algorithms and divergence considered (to the extent that our modeling of the neutral evolution of nucleotide sequence is appropriate). However, the accuracy of the evolutionary rate parameters appears to have little effect on accuracy, and the largest gain in accuracy and FPF is obtained from modeling the sequence content and the indel-length distribution. Parameterization of either is straightforward, so that our conclusions are relevant for practical alignment algorithms.

Posterior probabilities are a reliable estimator of alignment accuracy

In the previous section, we showed that posterior probabilities help to improve alignments. We next investigated whether they are also directly informative of alignment accuracy. Although posteriors cannot be used to distinguish correctly aligned columns from incorrect ones (except possibly when the posterior is either 0 or 1), they do provide a quantitative indication of reliability (Fig. 5C). In this section, we investigate the accuracy and robustness of this measure.

We calculated posteriors for all columns in simulated human–mouse sequences that were subsequently realigned using Viterbi decoding. Alignments columns were divided into 10 categories by their 10% posterior probability quantile. Within each category, we aggregated two statistics: the proportion of correctly aligned nucleotides, and the average percentage sequence identity. To test for robustness against modeling errors, this procedure was applied both for the basic and the full model. For both models, the posterior probability accurately predicts the proportion of correctly aligned columns (Fig. 8).

Similarly, the "asymptotic accuracy", defined as the proportion of correct alignment columns at distance 15 from the nearest gap, is very nearly identical to the average posterior at that distance, across a wide range of divergences (Fig. 3). Again, this conclusion remains true, even when the evolutionary model does not accurately fit the data (Supplemental Fig. S2).

Sequence identity and posterior probability show a strong, positive correlation. This is partly caused by an increasing admixture of nonhomologous nucleotide pairs as the posterior probability decreases. However, the observed PID for the highest posterior bin (>0.9) is 74.1%, exceeding the true PID of 69%. We interpret this as the result of stochastic effects that causes local sequence similarity to fluctuate, which in turn influences the accuracy with which alignments can be inferred. The result is that locally accurate alignments are biased toward regions with high-sequence identity. This suggests that it would be unwise to only use alignment columns with very high posterior probabilities to estimate substitution rates.

Comparison with score-based aligners

To put the performance of the probabilistic aligners in context, we realigned the simulated data using five general-purpose score-based aligners: ClustalW (Higgins and Sharp 1988), Lagan (Brudno et al. 2003), DiAlign (Morgenstern 1999, 2004), Mavid (Bray and Pachter 2004), and TBA/BLASTZ (Schwartz et al. 2003; Blanchette et al. 2004). The performance of these aligners was compared using the same three statistics as before (Fig. 9).

With the exception of DiAlign, all aligners achieve comparable sensitivities (79.4%–84.3%). BLASTZ paired this with good false-positive and nonhomologous fractions (FPF, 13.77%; NHF, 1.18%) when using the recommended score-threshold option (-K 2200). Lowering the score threshold to 2000 (following Pollard et al. 2004) increased the sensitivity from 79.4% to 82.0%, while the false-positive and nonhomology fractions increased only marginally (to 13.89% and 1.21%, respectively). The other score-based aligners were designed to perform global (or "glocal") (Brudno et al. 2003) alignment, thus solving a different problem that resulted in high (and less meaningful) NHF and FPF statistics. DiAlign was designed for multiple alignment of divergent protein-coding sequences, and as a consequence, is conservative in inferring homology, resulting in a low false-positive fraction

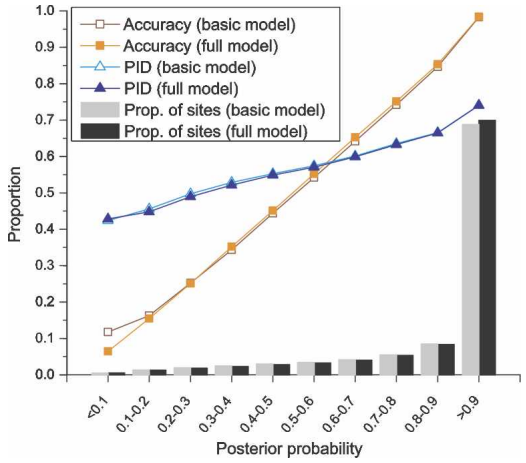


Figure 8. Posterior probability is an excellent indicator of alignment accuracy. Shown are the proportion of correctly aligned nucleotides (squares), the average sequence identity (triangles), and the proportion of nucleotides (histogram bars) across 10 posterior probability quantiles, obtained from realigned simulated human–mouse sequence data. For realignment, we used Viterbi decoding on the basic and full models.

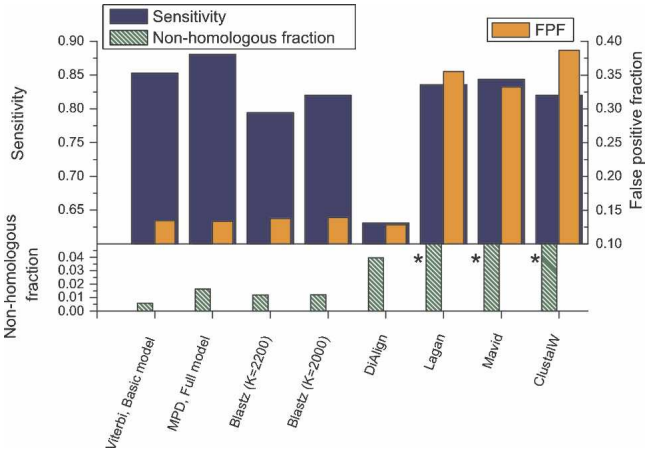


Figure 9. Performance comparison of score-based aligners. Histogram bars show sensitivity (black; *top left* axis), false-positive fraction (gray, *right* axis) and nonhomologous fraction (striped, *bottom left* axis), for simulated sequence based on human–mouse evolutionary parameters. The results for two probabilistic aligners (*leftmost* two sets) are included for comparison. Histogram bars marked by asterisks are off the scale; nonhomologous fraction for Lagan, 0.212; Mavid, 0.201; ClustalW, 0.223. Note that the axes in Figure 6 have different scales.

and a fair nonhomologous fraction (FPF, 12.8%; NHF, 3.95%), but a concomitant low sensitivity (63.1%). ClustalW was designed for protein multiple alignment, but was included because of its traditionally large user base. In our test, it shows lower sensitivity (81.9%) and higher false-positive rates (38.7%) than both Lagan and Mavid. However, despite their differences, all algorithms show qualitatively similar biases in their alignments (Supplemental Fig. S3), and uniformly do not perform as well as the MPD algorithm tested.

Discussion

In this study, we report on a large-scale simulation study, with the twofold aim of investigating the type and extent of biases that are inherent in the inference of alignments and of assessing whether a probabilistic approach can help reduce these biases.

We have distinguished three types of alignment biases; gap wander, gap attraction, and gap annihilation. Although well known, only one of these (gap wander or "edge wander") has, to our knowledge, been studied explicitly before (Holmes 1998). We have argued that gap wander is the dominant cause of wrongly aligned bases in maximum-likelihood alignments. This conclusion is supported by a theoretical analysis of gap wander under a Jukes–Cantor substitution model, the predictions of which agree very well with simulated data for small divergences. For higher divergences, additional biases start contributing to alignment inaccuracies, but gap wander continues to be important. For example, at a divergence of 0.375 substitutions and 0.05 indels per site, gap wander is predicted to cause 10% of homologous bases to be wrongly aligned. Simulations show the actual proportion to be 14%, the additional 4% apparently due to other biases.

These additional biases are caused by gap interactions, and their impact increases quadratically with divergence. The effects of gap attraction are apparent in the distribution of distances between successive gaps, in which small distances are strongly under-represented (Fig. 2B). Gap attraction strongly reduces the gap density in alignments, and further compounds the reduction of alignment accuracy near gaps that is caused by gap wander. A third and related bias, termed "gap annihilation", is also of second order in the divergence but occurs less frequently. In contrast to the other two biases, gap annihilation colocalizes with alignment gaps only very weakly (Lunter 2007), and causes an increase in both the apparent divergence and the error proportion across the alignment, and a further decrease in the number of inferred indels.

Both gap-interaction biases tend to decrease the alignment gap density compared with the true indel count. Increasing the indel rate of the inference model (i.e., lowering the gap-opening penalty) increases the number of inferred gaps, reducing this bias. However, our results show that the true evolutionary parameters do maximize the proportion of correctly aligned nucleotides, despite the gap count being negatively biased. In other words, the number of gaps can be made to approximate the true indel count, but only at the expense of placing the gaps in the wrong positions and increasing the proportion of incorrectly aligned bases.

It might seem that a tighter modeling of the evolutionary process would help to discern the true evolutionary history from among the many possibilities, and so reduce the impact of alignment biases. We found that more accurate modeling resulted in only very marginal improvements of the alignment accuracy. Indeed, in our simulation study of sequences at human–mouse divergence, the modeling of indel lengths using a mixed geometric distribution resulted in the single largest improvement in sensitivity, from 85.3% to 85.6% using Viterbi decoding, and from 87.8% to 88.2% using MPD. The geometric mixture model helps to align sequences across large indels, which are relatively infrequent, explaining the relatively modest improvement. Modeling the variation in GC content reduces the false-positive fraction (from 15.2% to 13.6% using MPD), but has little effect on sensitivity. Surprisingly, accurate modeling of indel and substitution rate variation has little, if any, effect. This robustness to misparameterization is supported by our simulations under the Jukes–Cantor model, where substantial variations in the rate parameters resulted in very little difference (Fig. 4).

For the data set of simulated sequences at human–mouse divergence, all models and decoding algorithms show 12%–15% wrongly aligned columns. This seems to reflect the loss of information during evolution rather than model inaccuracies or parameterization errors, and suggests that more sophisticated improvements to evolutionary models that might be considered, such as modeling evolving GC fractions (Lipatov et al. 2006), strand biases (Green et al. 2003), or context-dependent evolution (Jensen and Pedersen 2000; Arndt et al. 2003; Hwang and Green 2004; Lunter and Hein 2004; Siepel and Haussler 2004; Christensen 2006), although extremely valuable to help understand evolution, are unlikely to result in substantial improvement of sequence alignments. One aspect not modeled by any alignment algorithm that we are aware of is that indels often occur in tandem repetitive sequence as a result of, e.g., microsatellite instability (Kroutil and Kunkel 1999). The proposed mechanism, polymerase slippage, suggests that insertions often involve sequence duplications rather than insertions of random sequence. It would be interesting to investigate the possible improvement that modeling this aspect would have on alignment quality.

Modeling of polymerase slippage aside, we expect further improvements in alignments to arise chiefly from deep sequencing of extant species. Beside obvious factors such as the shape of the phylogenetic tree and the availability and quality of data

from genome-sequencing efforts, the achievable alignment quality will also, and probably crucially, depend on the quality of multiple alignment algorithms. Because the alignment problem suffers from a combinatorial explosion when the number of species increases, heuristic methods must be used. We have shown here that uncertainties in alignments are prevalent and unavoidable. Especially in multiple alignments, it is therefore essential that these uncertainties are dealt with properly. Many of the widely used multiple alignment algorithms "freeze" particular alignment choices at internal nodes, which would exacerbate alignment biases (Loytynoja and Goldman 2005), and more sophisticated methods than those currently available are required to optimally exploit the information that is available in multiple sequences.

Nevertheless, a simulation experiment showed that BLASTZ/TBA multiple alignments (Blanchette et al. 2004) do benefit from additional species (see Supplementary Information). We simulated sequences along the phylogeny of human, macaque, mouse, rat, and dog, and found that addition of in-group species consistently improved the implied human–mouse alignment (Supplemental Fig. S4). Adding all species resulted in an improvement of the sensitivity for human–mouse homology from 82% to 87.4%, similar to or slightly below the sensitivity of MPD pairwise alignments of human and mouse sequence alone. As pairwise alignments serve as input to TBA, the two approaches can conceivably be merged, and it would be interesting to investigate the improvements that the MPD algorithm can bring to TBA multiple alignments.

The score-based aligners we tested show similar kinds of biases to the probabilistic aligners, especially the Viterbi algorithm. This is not surprising, as these algorithms are formally very similar, and indeed, we found that the performance of the Viterbi aligner is similar to that of the best score-based aligner tested, BLASTZ. However, posterior decoding aligners, in particular MPD, have no score-based counterpart and perform better than both the Viterbi algorithm and BLASTZ in our simulations. MPD improves the sensitivity from 82% for BLASTZ to 88% for MPD, thus reducing the number of missed alignment columns by a third. Since our simulation procedure incorporated more aspects of human–mouse sequence evolution than any other that we are aware of, and was carefully parameterized using the best whole-genome human–mouse alignments currently available, it appears that the MPD algorithm would compute better alignments of mammalian genomic sequence than the current generation of score-based aligners is able to provide. Software to compute these alignments and a genome browser for human–mouse alignments are available at http://genserv.anat.ox.ac.uk/grape.

Comparing the score-based aligners among themselves, we observed large differences. The five score-based aligners we tested have each been designed with different purposes in mind, and the results reflect these design choices. For example, Mavid and Lagan are global aligners, and DiAlign was designed for aligning highly divergent sequence. The design aims of BLASTZ are the closest to our study, and it indeed performed the best out of the score-based aligners we tested.

Despite the performance differences between existing aligners and the scope for improvements of (particularly) multiple-alignment algorithms, our results suggest that alignment accuracy is fundamentally limited. For alignments of species at distances comparable to human and mouse, it seems likely that at least 10% of nucleotides in whole-genome alignments will remain wrongly aligned. These—unavoidable—errors need to be acknowledged and, where possible, accounted for when alignments are used to draw conclusions about evolution. We have shown that alignment uncertainties depend strongly on evolutionary distance, becoming less pronounced at lower divergences. However, alignment uncertainties are not spread uniformly over alignments, and sequence content (e.g., repetitive or near-repetitive sequence) also strongly influence the certainty with which alignments can be inferred, something that affects sequences at any divergence. We hope that the tools we provide will help researchers to identify and account for these local regions of uncertainty in alignments.

In conclusion, our results show that a probabilistic approach to sequence alignments has significant advantages over score-based approaches. Posterior probabilities are reliable and robust indicators of local alignment reliability. We have further shown that the MPD algorithm, and, to a lesser extent, improvements in evolutionary modeling, result in improvements in alignment quality. However, much uncertainty remains, and because of this, it seems inappropriate to continue the practice of using single most-likely alignments, essentially "point estimates without error bounds" (Zuker 1991). A probabilistic approach to sequence alignment is essential to properly account for and quantify these unavoidable uncertainties in alignments.

Methods

Definitions

Throughout this study, we scale units such that the divergence time between sequence pairs is 1, making the substitution rate σ equal to the expected number of substitutions per site. Because overlapping indel events are hard to deconvolute (Miklós et al. 2004), for convenience we here define the indel "rate," δ, as one minus the survival probability of a pair of homologous nucleotides, conditional on its left neighboring pair surviving. Equivalently, δ may be defined as the gap-opening probability in the true alignment. With the chosen units, this is numerically close to the indel event rate (Lunter 2007). We define the indel/substitution rate ratio as $\gamma = \sigma/\delta$.

To summarize the performances of aligners, we use the sensitivity (S), false-positive fraction (FPF), and nonhomologous fraction (NHF). S is defined as the ratio of correct alignment columns to all homologous columns. The FPF is defined as the proportion of wrongly aligned columns among all nongapped columns. We distinguish incorrect alignment columns involving "padding sequence" that does not share homology with other sequence, and all others. The proportion of columns containing padding sequence among all aligned columns summarizes the ability to tell alignable sequence from nonhomologous sequence, and is referred to as the nonhomologous fraction, NHF. Note that nucleotides contributing to NHF also contribute to the FPF (i.e., NHF < FPF). As a simple proxy for divergence, we use PID (proportion identity) throughout, most often as an aggregate measure (e.g., PID of nucleotides at particular distances from alignment gaps).

Evolutionary rates from human–mouse alignments

Substitution probabilities and other evolutionary parameters for the Full HMM model (see below and Fig. 5A) were obtained by maximum likelihood, using existing whole-genome BLASTZ human–mouse alignments as training data. Because all training data were considered to be homologous, we removed padding states from the model for training. Training data were stratified

according to the GC fraction (fGC), as measured in 250-bp windows and binned into 20 equally populated bins. Training was done separately for each fGC-bin (see Supplemental Table S2). Although our interest is primarily in alignment of bulk genomic, thus mainly neutral, DNA, we did not remove the small fraction of known functional or conserved sequence. Our stratification into fGC categories ensures that most protein-coding exons are found in the highest fGC category, while all other categories are dominated by neutrally evolving sequence.

To be able to model local substitution-rate variation (which does not strongly covary with fGC [Hellmann et al. 2005]), we estimated the spectrum of substitution rate by measuring local, putatively neutral sequence divergence on human–mouse ancestral repeats within 100-kb windows, across the human genome. These values were binned into 10 equally populated bins and averaged (see Supplemental Table S3), and used as input to the simulation stage (see below).

Models and probabilistic alignment

The probabilistic aligner consists of a standard three-state pair HMM, which was modified in two ways. First, we duplicated the insertion and deletion states to model the mixture-geometric indel length distribution. Second, we added four "padding" states, modeling the existence of nonhomologous sequence at either end of the alignable sequence. Parameters of the model are: δ, the gap opening probability; ε_1 and ε_2, the parameters determining the two geometric indel length distributions, α, their mixture coefficient, and τ, the alignment length parameter. The HMM topology is given in Figure 5A.

The parameter τ has a negligible effect on the alignments, and we fixed its value to 0.001. We defined six evolutionary models by restricting the ways in which the parameters vary with the input sequences. For the Full model, δ, ε_1, ε_2, and α all vary according to the fGC of the sequence. The substitution probabilities also depend on fGC and were additionally scaled to reflect the local sequence similarity. For the Basic model, all parameters were set to their average value, no scaling for local sequence similarity was done, and α was set to 1.0, corresponding to a standard geometric indel length distribution (i.e., affine-gap penalties). The remaining four models (see Table 1) resembled the Basic model, but each included one feature of the Full model: GCIndel, GCSubs, LocalSubs, and MixtureIndel included, respectively, fGC-dependent indel rates through δ; fGC-dependent substitution rates; local diversity-dependent substitution rates; and a mixture-geometric indel length distribution, which, however did not depend on fGC (α fixed at 0.857, see Supplemental Table S2).

Sequences were aligned using the Viterbi algorithm, and with two posterior decoding algorithms (see Appendix B). To reduce computation time, the dynamic programming tables were constrained by a banding procedure. The bandwidth was set independently of the sequences to be aligned by considering all paths corresponding to simulated alignments and computing the maximum deviation from the diagonal. To this maximum we added 15 to ensure that all sampled paths remained well inside the band, ensuring that the banding does not favorably bias the alignments.

Data simulation

Aligned sequence data were simulated by sampling from the Full model after removing the padding states. The simulated data set consisted of 100 sequence pairs for each of the 20 fGC categories and 10 substitution rate categories (20,000 pairs). After sampling, we padded each sequence with 100 nt of nonhomologous sequence at each end, drawn from the appropriate background distribution, resulting in sequences with an average length of 893 nucleotides (1153 alignment columns), of which 693 were alignable (753 columns), comprising in total 2×17.87 Mb.

For the Jukes–Cantor model, we sampled 500 sequence pairs for each of 16 evolutionary distances, ranging from $\sigma = 0.0$ to 0.9 in steps of 0.075, keeping $\gamma = 7.5$ throughout. Each sequence was padded with 100 bp of nonhomologous sequence at each end, resulting in an average of 872 nucleotides per sequence (1149 alignment columns), of which 672 were alignable (749 columns), comprising a total of 2×6.98 Mb.

Score-based alignments

We used the following score-based aligners: DiAlign 2.2 (Morgenstern 1999) with default parameters; Mavid 2.0.4 (Bray and Pachter 2004) with default parameters and a tree with total divergence 0.5; Lagan v1.21 (Brudno et al. 2003) with default parameters; ClustalW 1.83 (Higgins and Sharp 1988) with default parameters for DNA sequence. For BLASTZ version 7 (Schwartz et al. 2003) we used two parameter settings for the score threshold: K = 2200, which is recommended for human–mouse alignments, and K = 2000, below which alignments start to become less reliable (Pollard et al. 2004).

Acknowledgments

We thank Chris Ponting, Mikkel Schierup, and Michael Lässig for helpful discussions. G.L. and A.H. thank the MRC for financial support. A.R., N.M., and A.C. were supported by grants RAJLO (MRC), RHNIO (BBSRC), and RHJIHO (BBSRC) to J.H.

Appendix A: Analysis of gap wander

The maximum likelihood and true alignments need not be identical, because homology does not imply sequence identity and vice versa. Shifting a gap from its true location may thus increase the likelihood. Here we quantify this "gap wander" analytically for sequences evolving under a Jukes–Cantor model, following in part the analysis in Holmes (1998). We assume low indel rates, so that interactions between indels may be ignored.

Let L_0 be the log likelihood of the true alignment under the model of Figure 5A and L_i the likelihood of the alignment where a single gap is displaced rightward by i nucleotides from its true location. We first consider gap displacements in one direction only, assuming $i \geq 0$. Invoking the low indel rate assumption, shifting the gap by one nucleotide to the right does not cause collisions with other gaps, and so does not change their length or number. Consequently, any change in the log likelihood of the alignment is due to the replacement of a single alignment column containing homologous nucleotides (a "homologous column") by one containing nonhomologous nucleotides (a "nonhomologous column"). Both types of columns may contain either matching or nonmatching nucleotides. Under the Jukes–Cantor model, homologous columns contain matching and nonmatching nucleotides with probabilities $\frac{1}{4} + \frac{3}{4}e^{-4\sigma/3}$ and $\frac{3}{4} - \frac{3}{4}e^{-4\sigma/3}$, respectively. Shifting a gap by one nucleotide therefore causes the log likelihood of the alignment to increase or decrease by $S = \log(1 + 3e^{-4\sigma/3}/1 - e^{-4\sigma/3})$, or remain unchanged. The sequence of random variables $L_0, L_1, L_2, \ldots$ thus defines a random walk with steps $+S, 0, -S$. Denoting their probabilities by a,b,c and using that the probability of finding identical nucleotides in nonhomologous columns is $\frac{1}{4}$, we find $a = \frac{1}{4}(\frac{3}{4} - \frac{3}{4}e^{-4\sigma/3})$, $c = \frac{3}{4}(\frac{1}{4} + \frac{3}{4}e^{-4\sigma/3})$, and $b = 1 - a - c$. Since $a < c$, the random

walk $L_0,L_1,L_2,\ldots$ has negative drift, so that $M = \max_{i \geq 0} L_i$ exists with probability 1. Let T be the index for which this maximum is last reached, representing an optimal gap location (rightward of the origin). To derive the distribution $p_t = \Pr(T = t)$ of this location, we suppose a random walk $L_0,L_1,L_2,\ldots$ to be given, and construct another by adding one step in front. For the new random walk $L'_0,L'_1,L'_2,\ldots$ we have $L'_{k+1} = L_k$, and we denote the new maximum and index-of-last-maximum by M' and T'. We have $T' = T + 1$ and $M' = M$ unless $M = L_0$ and a step in the negative direction ($-S$) was added, since only in that case $M' = L_0 + S$ is the new maximum, and $T' = 0$. This implies that

$$p_{t+1} = p_t[1 - c \Pr(M = L_0)]. \quad (2)$$

In the language of random walks, the event $M = L_0$ is the escape probability of a random walk with drift and absorption and is computed as follows. Let q_k be the probability that the sequence $L_0,L_1,L_2,\ldots$ takes on the value 0 at least once ("absorption") when starting from $L_0 = kS$. For $k \geq 0$ we have $q_k = 1$ because of negative drift, while for $k < 0$ these probabilities satisfy $q_k = aq_{k+1} + bq_k + cq_{k-1}$, or $(q_{k+1} - q_k)/(q_k - q_{k-1}) = (c/a)$. Using the boundary conditions $q_0 = 1$ and $q_{-\infty} = 0$, this has the unique solution $q_k = (c/a)^k$, and in particular $\Pr(M = L_0) = 1 - q_{-1} = 1 - (a/c)$. Substitution into (2) yields $p_t = (c - a)(1 + a - c)^t = (1 - r)r^t$, where $r = 1 - \frac{3}{4}e^{-4\sigma/3}$. This describes the distribution of the maximum likelihood (ML) gap location rightward of the origin. The actual ML location is obtained by maximizing over locations both left and right of the true site. We approximate the deviation of the ML location away from the origin as $U = \max(T_L,T_R)$, where T_L,T_R are the left and right ML gap distances to the origin. This is an approximation, since the likelihood need not attain its maximum at the maximum distance; however, the conditional expectation of the maximum value given the distance is an increasing function of the distance, so the error introduced in this way is small. Using $\Pr(U \leq t) = \Pr(T \leq t)^2$, we find $\Pr(U = t) = (1 - r)r^t(2 - r^t - r^{t+1})$. The expected value of U is $E(U) = r(r + 2)/(1 - r)(1 + r)$, or

$$E(U) = \frac{(4e^{4\sigma/3} - 3)(4e^{4\sigma/3} - 1)}{8e^{4\sigma/3} - 3}, \quad (3)$$

representing the expected number of wrongly aligned nucleotides per gap due to gap wander. This number is nonzero even for $\sigma = 0$ because of possible homonucleotide runs (or, more generally, tandem repeats), which ambiguate gap placement even for sequences that are identical except for gaps. The gap density as a proportion of aligned sequence is (σ/γ), again ignoring interactions between gaps. Multiplying (3) by this fraction finally yields (1).

Appendix B: Posterior decoding algorithms

We used a variant of posterior decoding which computes the alignment that maximizes the cumulative log posterior probability of all columns that contribute to the alignment. Let $a_1 \ldots a_n$ and $d_1 \ldots d_m$ be "ancestor" and "descendant" sequences, and suppose M_{ij} is the posterior probability of aligning nucleotides a_i and d_j, which is identical to the posterior probability of being in a "match" state at position (i,j) in the dynamic programming table. Similarly, let D_{ij} be the posterior probability that a_i was involved in a deletion between the descendant's nucleotides d_j and d_{j+1}, and let I_{ij} denote the same for an insertion of d_j between the a_i and a_{i+1}. These posteriors were calculated from the dynamic programming tables of the standard Forward and Backward algorithms. In our model, two sets of four states correspond to insertions and deletions (two in the main alignment HMM, and two padding states). Because these states are mutually exclusive, to compute the posterior probabilities D_{ij} and I_{ij} we aggregate the posteriors for the relevant contributing states. The maximum total product of posteriors along an alignment path is computed by dynamic programming as follows:

$P_{00} \leftarrow 1$

For i from 0 to n:

 For j from 0 to m:

  $P_{ij} \leftarrow \max(P_{i-1,j-1}M_{ij},P_{i-1,j}D_{ij},P_{i,j-1}I_{ij})$,

where all references to indices out of bounds are regarded as being 0. After populating the array, P_{nm} contains the maximum total posterior. Finally, a traceback algorithm is used to find the corresponding posterior decoding path.

The MPD algorithm differs in the way gaps are treated. In the standard variant above, the posterior probabilities M_{ij}, D_{ij}, and I_{ij} measure the probabilities that particular HMM states are visited, conditional on the sequence data. The posterior for a nucleotide to align to a gap character distinguishes gaps based on their location in the secondary sequence, since such gaps are represented by different states in the dynamic programming table. This results in relatively low posteriors for gapped nucleotides, exacerbating gap-interaction effects. To counter this, the MPD algorithm marginalizes over all possible gap locations within the secondary sequence, replacing D_{ij} and I_{ij} by their marginalized counterparts, $D'_{ij} = \sum_{k=0}^{m} D_{ik}$, and $I'_{ij} = \sum_{k=0}^{n} I_{kj}$. The resulting posterior is interpreted as the probability that a particular nucleotide is unaligned, without specifying the precise location of the gap to which it contributes. Note that each path contributes to at most one of $D_{i0}, \ldots, D_{im}$ so that $D'_{ij} \leq 1$, and similarly for the insert states.

References

Altschul, S.F. and Erickson, B.W. 1986. Locally optimal subalignments using nonlinear similarity functions. *Bull. Math. Biol.* **48:** 633–660.

Arndt, P.F., Burge, C.B., and Hwa, T. 2003. DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.* **10:** 313–322.

Batzoglou, S. 2005. The many faces of sequence alignment. *Brief Bioinform.* **6:** 6–22.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Bray, N. and Pachter, L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14:** 693–699.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13:** 721–731.

Brudno, M., Poliakov, A., Salamov, A., Cooper, G.M., Sidow, A., Rubin, E.M., Solovyev, V., Batzoglou, S., and Dubchak, I. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.* **14:** 685–692.

Byers, T.M. and Waterman, M.S. 1984. Determining all optimal and near-optimal solutions when solving shortest path problems by dynamic programming. *Oper. Res.* **32:** 1381–1384.

Chao, K.M., Hardison, R.C., and Miller, W. 1993. Locating well-conserved regions within a pairwise alignment. *Comput. Appl. Biosci.* **9:** 387–396.

Chiaromonte, F., Yap, V.B., and Miller, W. 2002. Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.* **7:** 115–126.

Christensen, O.F. 2006. Pseudo-likelihood for non-reversible nucleotide substitution models with neighbour dependent rates. *Stat. Appl.*

Genet. Mol. Biol. **5:** Article 18. http://www.bepress.com/sagmb/vol5/iss1/art18.
Dewey, C.N. and L. Pachter. 2006. Evolution at the nucleotide level: The problem of multiple whole-genome alignment. *Hum. Mol. Genet.* **15 Spec No 1:** R51–R56. doi: 10.1093/hmg/dd1056.
Dewey, C.N., Huggins, P.M., Woods, K., Sturmfels, B., and Pachter, L. 2006. Parametric alignment of *Drosophila* genomes. *PLoS Comput. Biol.* **2:** e73. doi: 10.1371/journal.pcbi.0020073.
Ding, Y., Chan, C.Y., and Lawrence, C.E. 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11:** 1157–1166.
Do, C.B., Mahabhashyam, M.S., Brudno, M., and Batzoglou, S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15:** 330–340.
Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis*. Cambridge University Press, Cambridge, UK.
Edgar, R.C. and Batzoglou, S. 2006. Multiple sequence alignment. *Curr. Opin. Struct. Biol.* **16:** 368–373.
Elofsson, A. 2002. A study on how to best align protein sequences. *Proteins: Struct. Funct. Genet.* **46:** 300–309. doi: 10.1002/prot.10043.
Fariselli, P., Martelli, P.L., and Casadio, R. 2005. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics* (Suppl 4) **6:** S12. doi: 10.1186/1471-2105-6-S4-S12.
Goad, W.B. and Kanehisa, M.I. 1982. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucl. Acids Res.* **10:** 247–263. doi: 10.1093/nar/10.1.247.
Green, P., Ewing, B., Miller, W., Thomas, P.J., and Green, E.D. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33:** 514–517.
Gusfield, D., Balasubramanian, K., and Naor, D. 1994. Parametric optimization of sequence alignment. *Algorithmica* **12:** 312–326.
Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22:** 160–174.
Hellmann, I., Prufer, K., Ji, H., Zody, M.C., Paabo, S., and Ptak, S.E. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* **15:** 1222–1231.
Higgins, D.G. and Sharp, P.M. 1988. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73:** 237–244.
Holmes, I. 1998. "*Studies in probabilistic sequence alignment and evolution.*" Ph.D. thesis, University of Cambridge and The Sanger Centre, Cambridge, UK.
Holmes, I. and Durbin, R. 1998. Dynamic programming alignment accuracy. *J. Comput. Biol.* **5:** 493–504.
Hwang, D.G. and Green, P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci.* **101:** 13994–14001.
Jensen, J.L. and Pedersen, A.-M.K. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob.* **32:** 499–517.
Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H.N. Munro), pp. 21–132. Academic Press, New York.
Kall, L., Krogh, A., and Sonnhammer, E.L. 2005. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* (Suppl 1) **21:** i251–i257. doi: 10.1093/bioinformatics/bti1014.
Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 179–186.
Kroutil, L.C. and Kunkel, T.A. 1999. Deletion errors generated during replication of CAG repeats. *Nucl. Acids Res.* **27:** 3481–3486. doi: 10.1093/nar/27.17.3481.
Lassmann, T. and Sonnhammer, E.L. 2005. Automatic assessment of alignment quality. *Nucl. Acids Res.* **33:** 7120–7128. doi: 10.1093/nar/gki1020.
Lipatov, M., Arndt, P.F., Hwa, T., and Petrov, D.A. 2006. A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution. *J. Mol. Evol.* **62:** 168–175.
Loytynoja, A. and Goldman, N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci.* **102:** 10557–10562.
Lunter, G. 2007. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics* **23:** i289–i296. doi: 10.1093/bioinformatics/btm185.
Lunter, G. and Hein, J. 2004. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* (Suppl 1) **20:** I216–I223.doi: 10.1093/bioinformatics/bth901.
Lunter, G., Miklós, I., Drummond, A., Jensen, J.L., and Hein, J. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6:** 83.
Lunter, G., Ponting, C.P., and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* **2:** e5. doi: 10.1371/journal.pcbi.0020005.
Metzler, D. 2003. Statistical alignment based on fragment insertion and deletion models. *Bioinformatics* **19:** 490–499.
Meunier, J. and Duret, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21:** 984–990.
Mevissen, H.T. and Vingron, M. 1996. Quantifying the local reliability of a sequence alignment. *Protein Eng.* **9:** 127–132.
Miklós, I., Lunter, G.A., and Holmes, I. 2004. A "Long Indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.* **21:** 529–540.
Morgenstern, B. 1999. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15:** 211–218.
Morgenstern, B. 2004. DIALIGN: Multiple DNA and protein sequence alignment at BiBiServ. *Nucl. Acids Res.* **32:** W33–W36. doi: 10.1093/nar/gkh373.
Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.
Paten, B. and Birney, E. 2007. PECAN. http://www.ebi.ac.uk/~bjp/pecan.
Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., and Eisen, M.B. 2004. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5:** 6.
Prakash, A. and Tompa, M. 2007. Measuring the accuracy of genome-size multiple alignments. *Genome Biol.* **8:** R124. doi: 10.1186/gb-2007-8-6-r124.
Roshan, U. and Livesay, D.R. 2006. Probalign: Multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22:** 2715–2721.
Schlosshauer, M. and Ohlsson, M. 2002. A novel approach to local reliability of sequence alignments. *Bioinformatics* **18:** 847–854.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Sellers, P.H. 1979. Pattern recognition in genetic sequences. *Proc. Natl. Acad. Sci.* **76:** 3041.
Siepel, A. and Haussler, D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21:** 468–488.
Sun, Y. and Buhler, J. 2006. Choosing the best heuristic for seeded alignment of DNA sequences. *BMC Bioinformatics* **7:** 133.
Thorne, J.L., Kishino, H., and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33:** 114–124.
Tramontano, A., Leplae, R., and Morea, V. 2001. Analysis and assessment of comparative modeling predictions in CASP4. *Proteins* **5:** 22–38.
Tress, M.L., Jones, D., and Valencia, A. 2003. Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.* **330:** 705–718.
Waterman, M.S. 1983. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc. Natl. Acad. Sci.* **80:** 3123–3124.
Waterman, M.S., Eggert, M., and Lander, E. 1992. Parametric sequence comparisons. *Proc. Natl. Acad. Sci.* **89:** 6090–6093.
Waterston, R.H.K., Lindblad-Toh, E., Birney, J., Rogers, J.F., Abril, P., Agarwal, R., Agarwala, R., Ainscough, M., Alexandersson, P., An, S.E., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Zuker, M. 1991. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J. Mol. Biol.* **221:** 403–420.

Received May 21, 2007; accepted in revised form October 3, 2007.