Finding the balance between the mathematical and biological optima in multiple sequence alignment

Maria Anisimova,[1,2] Gina M. Cannarozzi,[1,2] David A. Liberles[3]

[1]Department of Computational Science, Swiss Federal Institute of Technology-Zurich, Switzerland

[2]Swiss Institute of Bioinformatics (SIB), Switzerland

[3]Department of Molecular Biology, University of Wyoming, Laramie, WY, USA

Abstract

Recent advances in evolutionary modelling and alignment methodology enable alignment of sequences with special features and incorporate structural and functional information. However, our reviewing experience and a recent study by Morrison[1] suggest that these newer methods are under-utilized (especially in the communities of molecular systematics and experimental biology), and the resulting alignments are often curated manually. Most often, no clear biological reasoning is invoked during manual alignment; instead only aesthetic qualities are considered, as measured by eye. Such subjectivity is not consistent with core scientific principles. Although we recognize that methodological problems still exist, computerized alignment methods are currently more realistic and can model a variety of evolutionary mechanisms. We also suggest future directions for the further improvement of automatic alignment methods based upon disconnects of existing methods with underlying biological mechanisms.

Alignment as a statement of homology

Multiple sequence alignment (MSA) is central to standard bioinformatics pipelines for comparative genomic and systematic analysis. An alignment is ultimately a statement of homology, so that each column in the alignment is thought to have descended from a common ancestral state in evolutionary history. This intimately links MSA to other downstream methods of evolutionary inference, including phylogenetic tree construction.

Evolutionary history and shared common origins of columns in an MSA are assessed by similarity of characters of sequence features. Indeed, dating back to Zuckerkandl and Pauling[2] and the origins of molecular clocks, it was realized that sequences diverge with time and that sequence identity or similarity was an indicator of sequence homology. From this, there is an expectation that sequences of increasing evolutionary distance will be increasingly divergent. The rate and patterns of divergence will be dictated by macro-evolutionary and population genetic processes and protein function and structure.[3] A recent study quantifying the relationship between rates of sequence and structural divergence suggested that structure diverges three to ten times more slowly than sequence, as aspects of protein structure are critical to protein function and are thus under strong negative selective pressure.[4,5] These considerations have led to the development of alignment methods based upon patterns of sequence divergence. While the first alignment algorithms were tuned to align primarily protein data[6,7] alignment algorithms have recently improved, diversified and became better adapted to data other than proteins, including DNA,[8-11] coding DNA,[12-16] RNA,[17,18] and to sequences with special features, such as repeats, rearrangements, and promoter regions.[19,20] Further advances in methodology provided bioinformaticians with more data on structural and functional features of proteins – and statistical methods were proposed to incorporate these features in alignment optimization algorithms.[21-25] Moreover, recent advances allow quality evaluation for each column aligned, providing valuable information for upstream analyses.[26-29]

Sadly, our reviewer experience and a recent survey[1] suggest that state-of-art alignment methods are not commonly used, but older and less accurate algorithms and their implementations are still preferred. Many empirical scientists recognize that the computational methods for sequence alignment are problematic, and often use manual alignment to adjust the alignment *by eye*. Overall less than 1% of surveyed papers used the best performing methods like MAFFT,[30] MUSCLE[31] and ProbCons,[26] while 50-75% resorted to the familiar CLUSTAL,[7] and manual intervention.[1] Here, we discourage manual editing of alignments, on the basis of its lack of objectivity and reproducibility. We urge the greater use of recent alignment techniques, including those that incorporate *a priori* knowledge where it is available. We recognize the importance of optimality criteria and suggest that manual alignments should be compared objectively. A better understanding of recent methodological advances and benchmarking will facilitate the use of better alignment methods. The prevalence of the first generation program CLUSTAL may be partially due to the fact that it is embedded in many web servers.

How do alignment methods compare?

One example of alignment is provided in Figure 1. Many proteins involved in signal transduction contain the Src-homologous SH3 domain of about 60 amino acids long. To demonstrate the range of the differences between alignment methods, a collection of seven SH3 domain sequences were aligned using six programs. Figure 1A shows the alignments (PDB identifiers: 2o9s-A, 1shg, 1gfc, 1pkt, 1srm, 1pnj, 2hsp). All sequences share a core beta-barrel structure consisting of five or six strands (yellow) arranged in two beta sheets. The structures 1pnj and 1pkt also contain an insertion with some helical structure (green).

The first alignment (Figure 1A) is a structural alignment from the program TM-Align[32] as implemented in STRAP.[33] There are several points worth remarking on the structural alignment. The sum of pairs score for the structural alignment is negative, reflecting the differences in alignment that one gets with a structural vs. a similarity scoring criterion. Additionally, the structural assignments for identical amino acids are different in 1pkt and 1pnj. As a consequence the identical string 'KGSLVAL' in the insertion is not aligned. This could be an example of the difference between structural and sequence homology, the phenomenon of *structure sliding along the sequence*, where non-homologous positions adopt structurally identical roles whereas the homologous positions play alternative roles in stabilizing structure.

Also shown are the alignments created with

Correspondence: Gina M. Cannarozzi, Department of Computational Science, Swiss Federal Institute of Technology-Zurich, 8092 Zurich, Switzerland.
E-mail: gina@cannarozzi.com

Key words: multiple sequence alignment, insertion and deletion, phylogeny, models.

Received for publication: 18 June 2010.
Revision received: 2 November 2010.
Accepted for publication: 2 November 2010.

Contributions: all authors contributed equally to this work.

Acknowledgments: MA and GMC are supported by the Swiss Federal Institute of Technology (ETH Zurich); MA also receives funding from the SNSF (award 31003A_127325); DAL receives funding from an NIH-INBRE grant to University of Wyoming and NSF award DBI-0743374.

Trends in Evolutionary Biology 2010; 2:e7
doi:10.4081/eb.2010.e7

default settings of Clustal W, PRANK, MUSCLE, T-COFFEE and MAFFT as implemented in STRAP and the FSA alignment from the FSA webserver. The differences in the alignments in the gap regions are noteworthy. T-Coffee and FSA create alignments with no gaps in the secondary structure. Thus the highest scoring algorithms also perform the best when aligning secondary structure (even without imposing structural constraints to guide the alignment process).

Figures 1B and 1C show the reliability assessment of M-COFFEE and FSA, obtained from their respective webservers. Both of these programs provide a color scale to assess reliability from blue (low confidence in homology) to red (high confidence in homology). Judging from the differences in the alignments, the probabilistic reliability assessment of FSA seems to provide a better view of the variability of the alignment.

Benchmarking is used to perform automated assessment of the performance of different alignment algorithms. Various benchmarking datasets have been constructed to standardize evaluation of protein alignment algorithms: BaliBASE,[34] HOMSTRAD,[35] PREFAB,[31] and OXBench[36] are based on 3D structural superposition. BaliBASE was the first large test set of protein families and is comprised of manually curated alignments, sorted into reference sets based on sequence identity, length and sequence characteristics. To avoid uncertainty in the test set, only core blocks of reliable aligned sequences are part of the reference alignments. Although this increases the reliability of the assignments, it avoids precisely the parts of the alignment that are difficult to align. Because of this problem, some benchmarking datasets (BaliBASE and OXBench) now include full length sequences as well. There are also test data sets based on simulated data (IRMBASE).[37,38] The benefit of simulations is that the correct alignment is known with no uncertainty. However, the relevance of the simulated data to real sequence evolution generates uncertainty of another kind. A recent development is the use of a phylogeny criterion rather than structure-based criteria to assess alignment algorithm performance.[39] In any case, the downstream use of the MSA as well as the properties of the sequence data may dictate the best performing algorithm.

The reality and the choice of the multiple sequence alignment optimality criterion

The search for optimal alignment using dynamic programming scales as a function of alignment length N and the number of sequences S (typically of $O(N^S)$ complexity), and is intractable, especially for MSAs in the genome-sequencing era. As a result, approximate algorithms or heuristic methods have been developed to rapidly align larger numbers of sequences. Of note are two classes of algorithms, divide and conquer and progressive alignments. Divide and conquer[40] and related methods like POA[41] subdivide the sequence into shorter fragments that are aligned and then combined to generate a global alignment. Progressive methods are the most prevalent and align sequences as a combination of pairwise alignments weighted by an underlying guide tree. The tree is traversed from leaves to

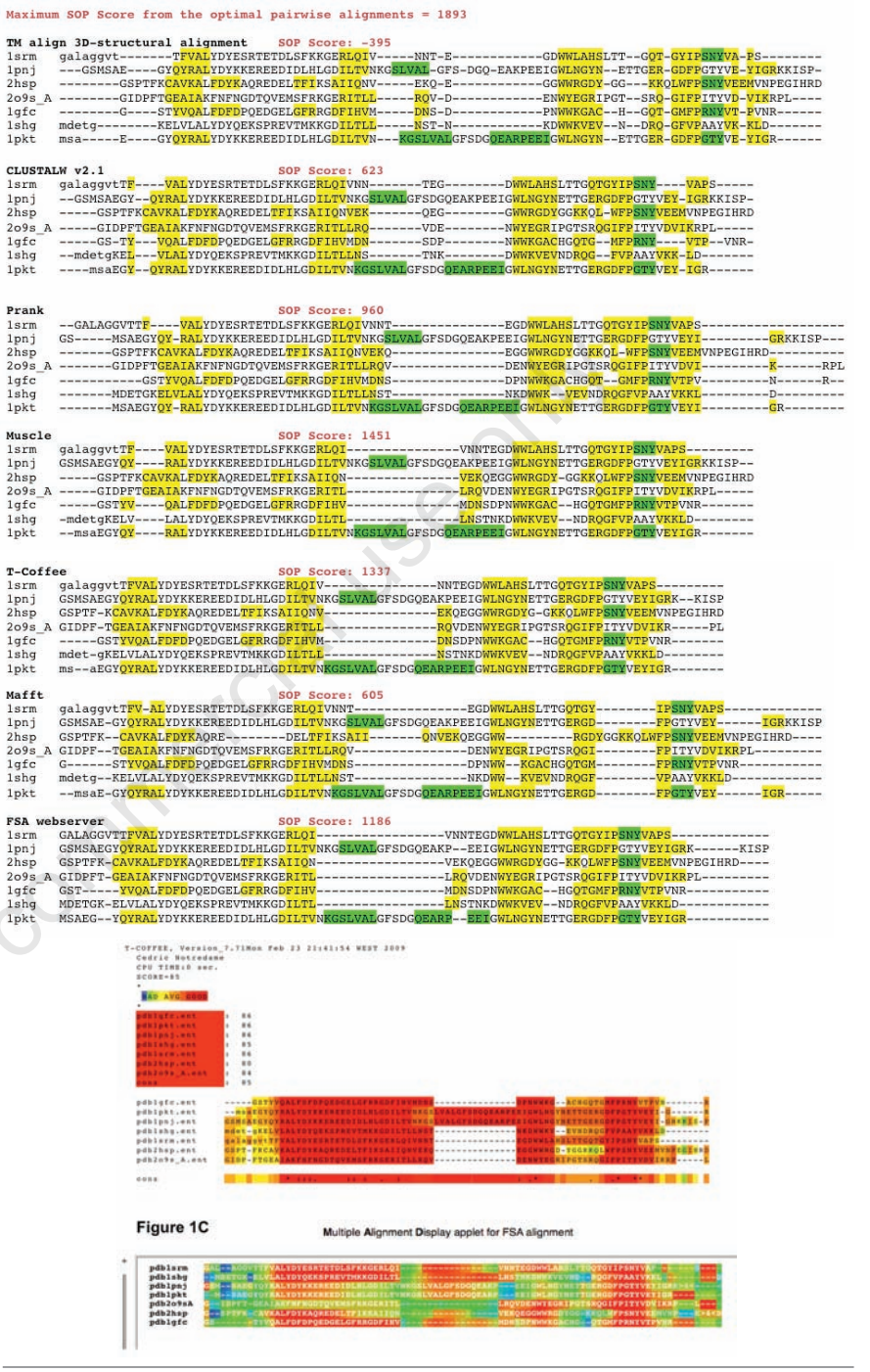

Figure 1. (A) Alignments of SH3 domains with PDB identifiers: 2o9s-A, 1shg, 1gfc, 1pkt, 1srm, 1pnj, and 2hsp.The alignments were performed with a structural aligner, TM-Align and a variety of alignment programs (Clustal W 2.1, PRANK, MUSCLE, T-Coffee, MAFFT, FSA). Beta strands are yellow while alpha-helices are green. (B) Reliability assessment of M-COFFEE. M-COFFEE provides reliability assessment as indicated by color on a scale from blue (low confidence in homology) to red (high confidence in homology). (C) Reliability assessment of FSA. FSA provides reliability assessment as indicated by color on a scale from blue (low confidence in homology) to red (high confidence in homology).

root (post-order traversal of the tree), aligning the more closely related sequences first and continuing by using the information at the child nodes to construct the partial alignment at the parent nodes. Alignment methods may differ in the manner in which the two sub-alignments at child nodes are combined. Common programs like Clustal,[7] MUSCLE,[31] and MAFFT[30] use progressive alignment as the underlying heuristic. More successful programs include an additional phase of iteration to correct mistakes created early in the progressive algorithm.

A critical part of all strategies is assessing the quality of the alignment, typically based on an optimality criterion that uses a character substitution or similarity matrix and a gap penalty scheme. The accuracy of an alignment method inherently depends on its optimality criterion, as it determines whether the alignment changes made by the heuristics may be accepted or not. Good optimality criteria are therefore critical. It has been pointed out eloquently by Kemena and Notredame[42] that the design of an optimality criterion should be motivated by the biological processes responsible for generating the molecular data at hand. Morrison[1] argued that the mathematical functions we optimize cannot ensure the convergence to the biological optimum as many underlying processes and features of real data are not adequately modelled, such as in the case of evolutionary scenarios with repeats, inversions, and frequent gaps. This, together with poor understanding of recent methodologies, leads researchers to resort to manual alignment.

Despite the inherent difficulties of modelling biological reality (no model is likely to ever be true), much progress has been made to improve alignment strategies and to make optimality criteria more realistic. Compared to a decade ago, we are now much closer to reaching the biological optimum, based on rigorous statistical criteria incorporating functional, structural or evolutionary views on alignment.[42] Relatively recently, pair hidden Markov models (HMMs) were applied to MSAs. Thanks to posterior decoding, where the most likely hidden state when an emission occurs is selected, HMM-based methods may incorporate more complex alignment scoring schemes.[26,27]

We note that defining the biological optimum may be very challenging. Scientists with different backgrounds (biochemists, evolutionary biologists, structural biologists, biophysicists, population geneticists, etc.) are likely to define distinct and non-overlapping biological optima. Ultimately, however, they are all describing biological processes that reflect descent from a common ancestor.

Indeed, such a discussion has already been broached with the suggestion that the best alignment for optimizing the evolutionary signal (phylogeny-driven) may not be the same as the best alignment based upon structural criteria. The use of structural criteria as benchmarks in multiple sequence alignment may then bias towards alignments optimized for structure and energetics, but which may not always reflect evolutionary history to the extent that sequences can slide through structures during evolution to find alternative thermodynamic optima.[39] The complex relationship between protein thermodynamics and population and macro-evolutionary processes including selection will need to be considered in generating an optimal scoring function as structural alignments will contain evolutionary information that is lost at the sequence level.[43]

Parsimony, homoplasies, and alignment

Manual editing of a computed MSA typically involves an even further minimization of the number of gaps and mismatches by eye. This is usually justified by a classical parsimony argument, but such practices disregard the underlying biological process, which may involve high degrees of complexity. While the computed alignments often are not the most parsimonious, evolution may not have generated the most parsimonious patterns of sequence diversity in aligned blocks.[3,44] Treating homoplasies as alignment mistakes excludes the possibility that a complex underlying process may have generated this pattern. To describe the inherent complexity of observed data, even parsimony methods attempt to include extra layers of complexity.[45]

Note that parsimony is not inconsistent with the presence of homoplasies. The distribution of observed patterns is defined by the underlying process, which generates data. An alignment with frequent gaps and homoplasies may well be the simplest description of data given the heterogeneous process resulting from an interplay of different evolutionary forces acting on molecular sequences (such as selection, recombination, compensatory changes, gene conversion, and composition biases in adaptation to the environment). It is a mistake to assume that aesthetically parsimonious patterns achieved by manual editing are representative of the true simplicity, without a better knowledge of the underlying biological process. Longer alignments are often more biologically meaningful than shorter and more parsimonious ones,[10,46] but would have been almost impossible to achieve by manual editing.

Three additional reasons for not trusting manually edited alignments

While it is hard to imagine accounting for complex biological forces in a manual alignment procedure, the alignment programs attempt to model the underlying biological process and use formal criteria to score the resulting alignments. While it is easy to underestimate the complexity of the alignment problem (especially for similar sequences), in computer science it has been long known as an NP-hard problem.[47] The likelihood of manually making "sensible" changes to an alignment rapidly decreases with the increase of number and the length of sequences to be aligned. But even for datasets of a manageable size, an important question is what we assume to be "sensible". The subjectivity of the manual alignment editing is one major pitfall that is inconsistent with doing proper science. Alignment algorithms use objective functions to evaluate and compare candidate alignments. Alignments obtained through subjective minimization of homoplasies without relying on an optimality criterion, are based on prejudices and open the way for introducing researcher-specific biases.

From this follows another important pitfall of manual alignment editing. The lack of statistical criteria to compare candidate alignments results in the inability to show that the manually edited alignment is significantly better than the one produced automatically. Finally, manual alignment curation is non-algorithmic and therefore not reproducible, defying one of the most important scientific criteria.

At the very least, accepting certain alignment alterations has to be done after: i) a statistical comparison of the manual alignment to other candidate alignments (for example, possible with M-coffee;[48] also[38,46,10] ii) the procedure for making the alignment changes has to be rigorously described and based on objective biological knowledge (for example, functional, active sites, structure elements) rather than on a parsimonious *gut feeling*.

One simple way of automatically curating alignment quality is to remove ambiguously aligned regions from subsequent phylogenetic analysis, which can be done with the popular program Gblocks.[49] However, the effect of applying Gblocks on downstream tree accuracy is controversial.[39,50] Strategically, if Gblocks is removing regions that are improperly aligned, then this is a band-aid covering the need for better models that produce better alignments. Alternatively, if Gblocks is removing properly aligned, but rapidly evolving regions, it is then introducing a bias to downstream analysis, as the most conserved sites may not be those that produce the strongest phylogenetic signal.[51]

The four disconnects in alignment methodology

Recent years have seen significant advances towards more biologically realistic alignments (see, for example, the review of Kemena and Notredame 2009).[42] Still, there

are several apparent discrepancies between the optimality criterion and the underlying biology; and a number of emerging independent trajectories were proposed in the literature to treat each such disconnect. The first disconnect involves the evolutionary process that presumably has contributed to generating the sequences in hand. However, optimality criteria used for scoring alignments are not based on evolutionary models, unlike the likelihood function in phylogenetic analyses that requires an explicit evolutionary model. This results in a disconnect: the alignment scoring function does not reflect the evolutionary process that contributed to generating the data. As better character substitution models are developed, they should also be applied to alignment optimization (*see discussion below*). However, our understanding of processes driving indel-formation is still poor, bringing us to the second disconnect: the lack of adequate indel models in alignment optimization. Gap penalties currently used by alignment methods are fundamentally different from the underlying probabilities of observing gaps of different lengths.[52] A uniform distribution of indels is usually assumed along the length of a sequence, which is also very unrealistic.[53-55] Modelling indel distributions and their evolution as part of an explicit evolutionary model is limited[56,57] and is currently not widely applied to alignments.

The third disconnect relates to the general lack of integration between population genetic and interspecific models. Indeed it has been shown that population genetics parameters like effective population size shape the interspecific patterns, and so affect the probability of observing different types of substitutions.[43,58-60]

Finally, the two competing views *functional or structural* vs *evolutionary* defining criteria for optimal alignment are currently disjoint: each alignment method chooses one criterion ignoring the other, and so creating a fourth disconnect. While the evolutionary view of alignment is most common, for proteins or RNA both indels and substitutions occur in the context of a folded three-dimensional structure, where their effects on $\Delta G_{folding}$ will affect their likelihood of observation, also dependent upon the unknown relationship between $\Delta G_{folding}$ and organismal fitness. Another problem rooted in the structural underpinnings of sequence evolution is the reliance of sequence-based alignment on substitution matrices with an underlying assumption of site-independence for a process that is inherently site-interdependent. Approaches that consider and combine sequence-based and structure-based criteria may be a promising step forward. What has not yet been accomplished is integration of ΔG calculations into sequence-based models for the purposes of alignment. With different underlying assumptions, this has been discussed in the context of fitness.[61-64]

Functional sites and sites that are critical to folding, such as active site and binding cleft residues that are absolutely conserved and cysteines responsible for structural disulphide bridges, can be pre-annotated and treated as anchors, providing a functional basis for alignment.

Better substitution models for an optimality criterion

Over the last decade much work has been done to improve models of molecular evolution. While BLOSUM[65] and PAM[66] matrices are still habitually used, it is time for the alignment methods to capitalize on the wealth of models available and include them in popular alignment packages. For protein-coding data the use of empirical codon matrices improves the alignment accuracy and has been implemented in alignment procedures.[10,12,14] For protein or RNA data, the use of more realistic similarity matrices should equally result in better alignments (eg, see RNA-specific matrix,[67] general protein matrix LG,[68] or organism-specific mtArt,[69] mtREV,[70] mtMam,[71] rtREV[72]). Due to among-site heterogeneity, structure or context-specific matrices have a strong potential to improve the alignment accuracy (for example, matrices for transmembrane alpha-helices,[73] for combinations of secondary structures and solvent accessibility,[74-77] or for local sequence-structure contexts[78]). These matrices give advantage in homology searches (especially when searching for distant homologues), but are poorly utilized for MSA. Biegert and Söding (2009) derived sequence context-specific amino acid similarities that rely on a library of sequence contexts, instead of relying on a single substitution matrix.[79] Based on this idea, a context-specific extension of BLAST (CS-BLAST) achieves a two-fold sensitivity improvement. The same idea potentially could help to improve alignment accuracy for distant sequences. In a related development, HMMs were successfully applied to describe and study the evolutionary heterogeneity of biological processes in a genomic sequence. For example, in application to G-protein-coupled receptors, models with hidden site classes were used to study the dimerization mechanisms.[80] Embedded within the phylogeny-aware alignment algorithm, a two-level HMM accounts for a number of heterogeneous classes describing distinct evolutionary processes, such as different codon positions or slow and fast evolving sites.[10]

Better indel models

The distribution of indels and their lengths is clearly dependent on several factors: sequence divergence, location within the sequence, proximity to other indels or functionally important regions, organism-specific factors. Insertion and deletion occur most commonly in loop regions of proteins, where they are less likely to cause steric problems (geometric clashes) in protein folding. Even in the absence of a solved structure, secondary structural information can be considered in alignment methods as different amino acids have different propensities to occur in loop regions.

While typical character substitution models assume a reversible and stationary Markov process at little price, for indels such assumptions are clearly unrealistic.[81,82] A simple model of affine gap penalties[83] is most frequently used, because few satisfactory and computationally tractable alternatives have been proposed. The choice of penalty parameters is rather arbitrary in practice (resorting to default values in most cases), although the choice of substitution matrix and gap penalties may be optimized.[84] While the indel length distribution is commonly described by a geometric (exponential) distribution, empirical studies suggested computationally more demanding solutions, such as the Zipfian distribution[85] or a mixture of four exponentials.[86]

Despite the difficult task, recent work on indel treatment for sequence alignment should not be underestimated. Pair-HMMs[6] and similar models such as transducers[87,88] were proposed to make modelling of indels more realistic. While Gotoh's gap penalties and the TKF[92] evolutionary model[57] may be equivalently modelled by a pair-HMM, the probabilistic framework allows additional sophisticated modifications to modelling indels. For example, the "long-indel" model is an extension of TKF[91] which allows indels of arbitrary length. The "long-indel" was shown to outperform both TKF models in sequence alignment and may be extended to a non-reversible process.[82] Alternative models (for example, exponential decay and extending the standard Markov process to include indel rates) were proposed to describe non-reversible time-dependent indel evolution and applied to gene finding.[89,81]

However, further methodological advances are required to make these recent models computationally feasible for MSA inference. In a more practical development, using more realistic bi-phasic gap penalties as in ProbCons[26] (gap-extension penalty is higher for shorter gaps) was shown to increase alignment accuracy. The most recent attempt to improve indel treatment in sequence alignment, with the aim of avoiding penalizing single insertion events multiple times, proposed to distinguish between insertions and deletions rather than treating them together. This 'phylogeny-aware' algorithm is implemented in the program PRANK, which relies on a tree to identify indel regions as insertions and deletions and treats them as such in the subsequent alignment

Table 1. Multiple sequence alignment software packages currently available to users (incomplete list).

Software package	Characteristic description	Indel model	Input data	Speed[1]	Uncertainty estimates
Standard alignment programs					
Clustal	The pioneering and most widely used program to construct MSAs.[7] Historically very important, but is outperformed by many current alignment programs. A recent version allows for iteration.[111]	Affine with position-specific penalties	AA, DNA	Fast	No
FSA	A fast probabilistic approach that seeks to minimize expected distance to true alignment (the expected accuracy objective function) and constructs MSA using sequence annealing based on pairwise estimates of homology.[87,27] FSA can align thousands of sequences, and also provides the capability to compare different candidate alignments provided by the user (serves as a meta-method).	Pair-HMM (equivalent to double affine penalties)	AA, DNA	Fast-very fast	Yes
T-Coffee	Coffee is an objective function for computing a consistency score based on the agreement between a library of pairwise alignments of the same sequences.[112] The library of alignments can also come from other alignment programs, structural information or be computed from the input sequences. T-coffee is an implementation using this objective function combined with a progressive alignment.[113]	No explicit penalty scheme (embedded algorithms typically use affine scheme). (like in ProbCons)	AA, DNA	Moderate	No
POA	Based on "partial order alignments", a representation of an MSA as a directed acyclic graph.[8,114] Partial orders can be aligned using dynamic programming using both progressive and iterative algorithms. POA includes the capability to model homologous recombination. Non-affine penalty scheme, which includes gap truncation penalty as well as the standard gap open and extension penalties.	T-coffee rewards correctly placed gaps. Recent version includes double affine penalty scheme	AA, DNA	Fast-very fast	No
PRANK	A "phylogeny aware" alignment program that uses a phylogenetic tree to recognize insertions and deletions as separate evolutionary events.[10,46] This algorithm was then extended to model regional heterogeneity and evolution.[9] This algorithm performed well in the performance analysis based on phylogeny.[39]	Pair-HMM (equivalent to affine), avoids repeatedly penalizing the same insertion by using a tree to distinguish it from deletion	DNA, AA, codons	Slow-moderate	Yes
MAFFT	Offers a variety of alignment strategies including progressive alignment (with very efficient dynamic programming algorithm), iteration, and consistency scoring and allows for choosing from a wide spectrum of accuracy and speed.[30] Mafft performs well on many benchmarks.[115,39]	Affine with position-specific penalties	AA, DNA	Very fast	No
MUSCLE	MUSCLE employs pairwise profile alignment with two steps of subsequent refinement.[94]	Affine with position-specific penalties	AA, DNA	Fast - Very fast	No
ProbAlign	Constructs MSAs by maximizing expected accuracy and using partition function methodology and a probabilistic consistency transformation scheme.[116]	Affine	AA, DNA, RNA	Moderate	Yes
ProbCons	Uses maximum expected accuracy algorithm, which combines probabilistic modeling and consistency-based scoring with a pair HMM-based progressive alignment algorithm.[26] This algorithm performs well on structural benchmarks and simulations.[115] A separate version exists for RNA (ProbConRNA).	Pair-HMM (equivalent to double affine penalty scheme)	AA	Moderate	Yes
MUMMALS	Uses a probabilistic consistency objective function and pairwise alignment with structural pair-HMM that considers local secondary structure similarities.[117]	Pair-HMM (equivalent to affine penalty scheme)	AA	Moderate	Yes
Template based alignment programs (using structure, profiles or other features)					
R-Coffee	Uses RNA structural template computed RNAlpfold and constructs an MSA having the best agreement of sequences and structures.[118]	No explicit penalty scheme.	RNA	Moderate	No
3D-Coffee	Combines sequences and structures to generate high-quality multiple sequence alignments.[119] EXPRESSO is a version of 3D-Coffee that automatically selects templates via a BLAST search against PDB.[120] T-Coffee may also combine MSAs with profile information (PSI-Coffee, 3DPSI-Coffee).	No explicit penalty scheme.	AA, DNA	Moderate	No

Continued next page.

Table 1. Continued from previous page.

Software package	Characteristic description	Indel model	Input data	Speed[1]	Uncertainty estimates
Praline	Homology extension package that uses a PSI-BLAST profile to guide the computation of an MSA.[121,122] PRALINE also provides a choice of secondary structure prediction programs that can be used for integrating structural information into the alignment process. Special option in PRALINE is tailored for membrane-bound proteins, by using prediction of transmembrane regions and membrane-specific scoring matrices.[123]	Affine	AA	Moderate	No
PROMALS3D	PROMALS3D uses 3D structural information to guide sequence alignments constructed by PROMALS,[22] which is a consistency-based aligner that uses libraries generated with pair-HMM posterior decoding strategy. Sequences in PROMALS are associated with a PSI-BLAST profile.				Yes??
Meta methods					
Guidance	Guidance score tests the consistency of every MSA column obtained from guide trees with respect to a set of MSAs, which is shown to be a good predictor of unreliably aligned regions.[29]	No explicit penalty scheme	DNA, AA, codons	Moderate (depends on embedded methods)	Yes, HoT or Guidance score
M-Coffee	A meta-method, which uses other alignment methods to construct an alignment library and then uses T-coffee to combine these alignments.[48]	No explicit penalty scheme	DNA, AA	Moderate (depends on embedded methods)	Yes, consensus-based
Joint estimation of alignment and phylogeny					
BAli-Phy	Joint Bayesian estimation employing Markov Chain Monte Carlo to sample trees, alignments and evolutionary model parameters.[92,96]	Pair-HMM (equivalent to affine gap penalties)	DNA, AA	Very slow	Yes
StatAlign	Joint Bayesian estimation employing Markov Chain Monte Carlo with fast transition kernels to sample trees, alignments and evolutionary model parameters.[97]	Modified TKF92	DNA, AA	Very slow	Yes
BigFoot	An extension to the StatAlign package, BigFoot performs phylogenetic footprinting by modeling quickly and slowly evolving regions and their breakpoints.[98]	Extension of TKF92 via HMM transducer	DNA	Very slow	Yes
ALIFRITZ	Employs a strategy based on simulated annealing to infer the tree, the alignment	TKF92 implemented via triplet HMM and the history of insertions and deletions.[99]	DNA, AA	Very slow-slow	No
Genomic aligners					
MLAGAN	A multiple genome aligner that uses progressive alignment guided by a user-specified tree and based on a sum of pairs metric.[106] It is based on LAGAN, a pairwise genome aligner which first maps local alignments to the genome and then uses the map in a global alignment phase. MLAGAN assumes a given species tree.	Affine	A set of DNA contigs	Fast	No
Enredo/Pecan	Enredo finds colinear segments of the input genomes and treats both genome rearrangements and duplications. It is used in conjunction with Pecan, which then aligns full genomes using a consistency objective function.[110] Pecan uses maximum expected accuracy criterion.	Pair-HMM (equivalent to double affine penalty scheme)	A set of chromosomes, each represented by linear or circular strings of double-stranded DNA	Moderate	Yes
Ortheus	A probabilistic approach that infers the evolutionary history of a multiple sequence alignment in terms of substitutions, insertions and deletions and so constructs an ancestral MSA.[110] Ortheus takes a regular MSA as an input but does not rely on a single fixed alignment.	Models indel history using a transducer	A phylogeny and an MSA	Moderate	Yes

[1]Summarized from evaluations in (87,39,124,110,125,108)

building process. PRANK has been shown to outperform implementations of other algorithms in analysis of biological data.[39] 'Phylogeny-aware' alignment has the potential to overcome the difficulties in aligning repeat regions identified as problematic by Morrison.[1]

Simultaneous estimation of alignment and phylogeny

Most alignment methods rely on a guide tree and thus may be affected by using a "wrong" tree to guide the alignment process. On the other hand, phylogenetic inference is typically conducted for a given alignment, and alignment errors may have a downstream effect on the accuracy of the inferred tree.[90] Methods have been developed that enable the use of evolutionary criteria in an iterative or simultaneous assessment of alignment together with the phylogenetic tree.[29,91-94] For example, the POY software[95] makes this assessment based upon a parsimony score. Statistically more rigorous developments include methods like BAliPhy,[92,96] StatAlign[97] and BigFoot,[98] which are formulated in a Bayesian framework, include models of indel evolution (TKF and extensions), and use simulated annealing[99] or MCMC to jointly sample posterior distributions of alignments and trees (eg, see Table 1). As a result, such methods are computationally demanding and currently cannot be used for large datasets.

Consequently, it appears tempting to use approximate schemes such as the iterative approach of Saté,[91] which at each step of the iteration attempts to improve the alignment and the tree using the best alignment algorithms and ML tree estimation by RAxML.[100] Such an approach however is controversial since the tree-building step has no indel model and thus will introduce biases to further iterative steps, especially for divergent sequences.

Overall, the success of alignment-phylogeny co-estimation relies not only on the ability to properly explore the joint tree-alignment space but also on the underlying models of character and indel evolution. However, these are similar models that are being used for the assessment of phylogeny, indicating the need for better models, especially in the placement of indels and integration of thermodynamic considerations. Improved models will then not only improve MSA inference, but also contribute to more accurate phylogenetic tree construction.

Manual alignment vs. the need for better models and methods

Here we suggested that manual alignment as commonly practiced suffers from being *ad hoc* and is often based upon faulty assumptions about the nature of evolutionary processes. This also has been emphasized in an appraisal of manual alignments by Giribet[101] who compared manually edited alignments from different manual-alignment experts. Indeed, the complexity of the alignment problem is easily underestimated when very similar sequences are aligned. But for low levels of divergence, the best alignment programs perform very well. For deeper divergences or larger samples, the prospect of 'successful' manual curation rapidly decreases, and in the genomics era, curated alignments rapidly lose their appeal. For "difficult" datasets (divergent, with long sequences or many taxa) current automatic approaches are unlikely to be outperformed by manual editing. Greater use of prior knowledge about the data (like 3D structure, knowledge of active sites, a known pattern expected for a protein domain, etc.) facilitates alignments of greater accuracy that cannot be achieved by hand, but has to be incorporated in the optimality criteria.

Clearly, better models and methods are needed (and appear to be on their way). In the meantime, a systematic assessment of assumptions and manual evaluation of the results of different approaches, including progressive and structural alignment with a clear criterion to integrate the two lines of information appears to be the best way forward. For large-scale approaches, this is clearly not possible and inference should be made with an awareness of any methodological weaknesses or faulty assumptions.

Concluding thoughts

Even the best MSA algorithms produce a certain alignment error, which rapidly increases with divergence. Concerns about alignment accuracy may be better treated if scientists understand that any alignment carries an element of uncertainty. For example, Aurahs *et al*.[102] is one of very few studies to consider several candidate MSAs to infer trees. The uncertainty in MSA inference has to be taken into consideration when making strong conclusions based on one alignment, as equally optimal alignments may lead to different inferences. This calls for a rigorous framework for comparison of candidate alignments, as it is often done with phylogenetic trees.[103] Simultaneous Bayesian inference of the alignment and tree provides one way of dealing with such uncertainty where a distribution of alignments and trees is the focus rather than single inferences, obtained in a frequentist framework.

The choice of alignment algorithms should be guided by the type of data, including its size, special features and availability of structural and functional information (Table 1). Applying several suitable alignment algorithms and then using a meta-method to evaluate candidate alignments currently is the best option for navigating through the vast space of possible alignments.

Better models that link the molecular and evolutionary mechanisms underlying the substitution and indel processes to MSAs are in their infancy, but the field is developing. Currently, inference can be made using existing computational approaches and thoughtful considerations of underlying assumptions.

It is exciting that alignment and underlying models are re-emerging as hot topics in bioinformatics. The renewed interest in alignment methodology is caused by growing demands for the analyses of large-scale and genomic data. The recent method FSA can align thousands of long sequences, while using a pair HMM to approximate the indel process on a tree and pairwise alignments based on the sequence annealing algorithm (Table 1).[27] Advances in sequencing technologies have allowed the rapid sequencing of full genomes, which in turn is driving advances in methodology for aligning and assembling short reads[104,105] and for multiple whole genome alignment (eg, see Table 1; MLAGAN,[106] Enredo and Pecan,[107-109] Ortheus[110]). Despite a number of recent algorithmic advances the genomics alignment field is still in its infancy, presenting succulent challenges, yet to be solved.

References

1. Morrison DA. Why would phylogeneticists ignore computerized sequence alignment? Syst Biol 2009;58:150-158.
2. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol 1965;8:357-66.
3. Liberles DA, Dittmar K. Characterizing gene family evolution. Biol Proced Online 2008;10:66-73.
4. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence - a study of structural response in protein cores. Proteins 2009;77:499-508.
5. Chothia C, Lesk AM. The evolution of protein structures. Cold Spring Harb Symp Quant Biol 1987;52399-405.
6. Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press;1998.
7. Thompson JD, Higgins DG, Gibson TJ. Improved sensitivity of profile searches through the use of sequence weights and gap excision. Comput Appl Biosci 1994;10:19-29.
8. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics 2002;18:452-64.
9. Löytynoja A, Goldman N. A model of evolution and structure for multiple sequence alignment. Philos Trans R Soc Lond B Biol Sci 2008;363:3913-9.
10. Löytynoja A, Goldman N. Phylogeny-aware

gap placement prevents errors in sequence alignment and evolutionary analysis. Science 2008;320:1632-5.
11. Kim K, Kim M, Woo Y. A DNA sequence alignment algorithm using quality information and a fuzzy inference method. Prog Nat Sci 2008;18:595-602.
12. Schneider A, Cannarozzi GM, Gonnet GH. Empirical codon substitution matrix. BMC Bioinformatics 2005;6:134.
13. Doron-Faigenboim A, Pupko T. A combined empirical and mechanistic codon model. Mol Biol Evol 2007;24:388-97.
14. Anisimova M, Kosiol C. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 2009;26:255-71.
15. Kosiol C, Holmes I, Goldman N. An empirical codon model for protein sequence evolution. Mol Biol Evol 2007;24:1464-79.
16. Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol (Clifton, N.J.) 2009;537: 39-64.
17. Hofacker IL, Bernhart SHF, Stadler PF. Alignment of RNA base pairing probability matrices. Bioinformatics 2004;20:2222-7.
18. Notredame C, O'Brien EA, Higgins DG. RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res 1997;25: 4570-80.
19. Blanco E, Guigó R, Messeguer X. Multiple non-collinear TF-map alignments of promoter regions. BMC Bioinformatics 2007;8:138.
20. Phuong TM, Do CB, Edgar RC, Batzoglou S. Multiple alignment of protein sequences with repeats and rearrangements. Nucleic Acids Res 2006;34:5932-42.
21. Poirot O, Suhre K, Abergel C, et al. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Res 2004;32:W37-40.
22. Pei J, Grishin NV. MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res 2006;34:4364-74.
23. Subramanian AR, Hiran S, Steinkamp R, et al. DIALIGN-TX and multiple protein alignment using secondary structure information at GOBICS. Nucleic Acids Res 2010;38:1-4.
24. Kececioglu J, Kim E, Wheeler T. Aligning protein sequences with predicted secondary structure. J Comput Biol 2010;17:561-80.
25. Bremges A, Schirmer S, Giegerich R. Fine-tuning structural RNA alignments in the twilight zone. BMC Bioinformatics 2010;11:222.
26. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genom Res 2005;15:330-40.
27. Bradley RK, Roberts A, Smoot M, et al. Fast statistical alignment. PLoS Comput Biol 2009;5:e1000392.
28. Ahola V, Aittokallio T, Vihinen M, Uusipaikka E. A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics 2006;7:484.
29. Penn O, Privman E, Landan G, et al. An alignment confidence score capturing robustness to guide-tree uncertainty. Mol Biol Evol 2010;27:1759-1767.
30. Katoh K, Misawa K, Kuma Ki, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30: 3059-66.
31. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32: 1792-7.
32. Pandit SB, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics 2008;9: 531.
33. Gille C, Frömmel C. STRAP: editor for STRuctural Alignments of Proteins. Bioinformatics. 2001;17:377-8.
34. Bahr A, Thompson JD, Thierry JC, Poch O. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res 2001;29:323-6.
35. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein science. 1998;7: 2469-71.
36. Raghava GPS, Searle SMJ, Audley PC, et al. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics 2003;4:47.
37. Subramanian A, Weyer-Menkhoff J, Kaufmann M, Morgenstern B. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics. 2005;6: 66.
38. Lassmann T, Sonnhammer EL. Automatic extraction of reliable regions from multiple sequence alignments. BMC Bioinformatics 2007;8:S9.
39. Dessimoz C, Gil M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol 2010;11:R37.
40. Stoye J. Multiple sequence alignment with the divide-and-conquer method. Gene 1998;211GC45–56.
41. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics 2002;18:452-64.
42. Kemena C, Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 2009; 25:2455-65.
43. Huzurbazar S, Kolesov G, Massey SE, et al. Lineage-specific differences in the amino acid substitution process. J Mol Biol 2010;396:1410-21.
44. Rokas A, Carroll SB. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol 2008;25:1943-53.
45. Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet 2001;17:262-72.
46. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 2005;102:10557-62.
47. Wang L, Jiang T. On the complexity of multiple sequence alignment. J Comput Biol 1994;1:337-48.
48. Wallace I, O'Sullivan O, Higgins D. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 2006;4:1692-9.
49. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 2000;17:540-52.
50. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56: 564.
51. Massey SE, Churbanov A, Rastogi S, Liberles DA. Characterizing positive and negative selection and their phylogenetic effects. Gene 2008;418:22-6.
52. Chang MSS, Benner SA. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol 2004;341:617-31.
53. Ponting CP, Lunter G. Signatures of adaptive evolution within human non-coding sequence. Human Molecular Genetics 2006;15:R170-5.
54. Lunter G, Ponting C, Hein J. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol 2006;2:e5.
55. Creer S. Choosing and using introns in molecular phylogenetics. Evol Bioinform 2007;3:99-108.
56. Thorne J, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol 1991;33:114-24.
57. Thorne JL, Kishino H, Felsenstein J. Inching toward reality: an improved likelihood model of sequence evolution. J Mol

Evol 1992;34:3-16.
58. Thorne J, Choi S, Yu J, et al. Population genetics without intraspecific data. Mol Biol Evol 2007;24:1667-77.
59. Higgs PG. Compensatory neutral mutations and the evolution of RNA. Genetica 1998;102-103:91-101.
60. Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 2008;25:568-79.
61. Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins 2002;46:105-9.
62. DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 2005;6:678-87.
63. Rastogi S, Reuter N, Liberles DA. Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophys Chem 2006;124:134-44.
64. Robinson DM, Jones DT, Kishino H, et al. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 2003;20:1692-704.
65. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17:49-61.
66. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 1978;5:345-352.
67. Savill NJ, Hoyle DC, Higgs PG. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics 2001;157:399-411.
68. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol 2008;25:1307-20.
69. Abascal F, Posada D, Zardoya R. MtArt: a new model of amino acid replacement for Arthropoda. Mol Biol Evol 2007;24:1-5.
70. Adachi J, Hasegawa M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 1996;42:459-68.
71. Yang Z, Nielsen R, Hasegawa M. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 1998;15:1600-11.
72. Dimmic M, Rest J, Mindell D. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol 2002;55:65-73.
73. Muller T, Rahmann S, Rehmsmeier M. Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics 2001;17:S182-S189.
74. Rice D, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence1. J Mol Evol 1997;267:1026-38.
75. Gong S, Blundell TL. Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. PLoS Comput Biol 2008;4:e1000179.
76. Goonesekere NCW, Lee B. Context-specific amino acid substitution matrices and their use in the detection of protein homologs. Proteins 2008;71:910-9.
77. Koshi JM, Goldstein RA. Context-dependent optimal substitution matrices. Protein Eng 1995;8:641-5.
78. Huang YM, Bystroff C. Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. Bioinformatics 2006;22:413-22.
79. Biegert A, Söding J. Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A 2009;106:3770-5.
80. Soyer OS, Dimmic MW, Neubig RR, Goldstein RA. Dimerization in aminergic G-protein-coupled receptors: application of a hidden-site class model of evolution. Biochemistry 2003;42:14522-31.
81. Rivas E, Eddy SR. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput Biol 2008;4:e1000172.
82. Miklós I, Lunter GA, Holmes I. A "Long Indel" model for evolutionary sequence alignment. Mol Biol Evol 2004;21:529-40.
83. Gotoh O. An improved algorithm for matching biological sequences. J Mol Evol 1982;162;705-8.
84. Edgar RC, Batzoglou S. Multiple sequence alignment. Curr Opin Struct Biol 2006;16:368-73.
85. Benner SA, Cohen MA, Gonnet GH. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Evol 1993;229:1065-82.
86. Qian B, Goldstein RA. Distribution of Indel lengths. Proteins 2001;45:102-4.
87. Bradley RK, Holmes I. Transducers: an emerging probabilistic framework for modeling indels on trees. Bioinformatics 2007;23:3258-62.
88. Holmes I. Using guide trees to construct multiple-sequence evolutionary HMMs. Bioinformatics 2003;19:i147-57.
89. Rivas E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics 2005;6:63.
90. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science 2008;319:473-6.
91. Liu K, Raghavan S, Nelesen S, et al. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 2009;324:1561-4.
92. Redelings BD, Suchard MA. Joint Bayesian estimation of alignment and phylogeny. Syst Biol 2005;54:401-18.
93. Yue F, Shi J, Tang J. Simultaneous phylogeny reconstruction and multiple sequence alignment. BMC Bioinformatics 2009;10:S11.
94. Edgar RC, Sjölander K. SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics 2003;19:1404-11.
95. Varon A, Vinh LS, Wheeler WC. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 2010;26:72-85.
96. Suchard MA, Redelings BD. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 2006;22:2047-8.
97. Novák A, Miklós I, Lyngsø R, Hein J. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics. 2008;24:2403-4.
98. Satija R, Novák A, Miklós I, et al. BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. BMC Evol Biol 2009;9:217.
99. Fleissner R, Metzler D, von Haeseler A. Simultaneous statistical multiple alignment and phylogeny reconstruction. Systematic Biology 2005;54:548-61.
100. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22:2688-90.
101. Giribet G, Wheeler WC, Muona J. DNA multiple sequence alignments. EXS 2002;107-14.
102. Aurahs R, Göker M, Grimm GW, et al. Using the Multiple Analysis Approach to Reconstruct Phylogenetic Relationships among Planktonic Foraminifera from Highly Divergent and Length-polymorphic SSU rDNA Sequences. Bioinformatics and Biology Insights. 2009;3:155-77.
103. Aris-Brosou S. How Bayes tests of molecular phylogenies compare with frequentist approaches. Bioinformatics 2003;19:618-24.
104. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nature Biotechnology. 2009;27:455-7.
105. Chaisson MJ, Brinza D, Pevzner P. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res 2009;19:336-46.
106. Brudno M, Malde S, Poliakov A, et al. Glocal alignment: finding rearrangements during alignment. Bioinformatics 2003;19:i54-62.
107. Paten B, Herrero J, Beal K, et al. Enredo

and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome research. 2008; 18:1814-28.
108. Paten B, Herrero J, Beal K, Birney E. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics 2009;25:295-301.
109. Brudno M, Do CB, Cooper GM, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 2003;13:721-31.
110. Paten B, Herrero J, Fitzgerald S, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 2008;18:1829-43.
111. Larkin M, Blackshields G, Brown N, Chenna P. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947-8.
112. Notredame C, Holm L, Higgins D. COFFEE: an objective function for multiple sequence alignments. Bioinformatics 1998;14:407-22.
113. Notredame C, Higgins D, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment1. J Mol Biol 2000;8:205-17.
114. Grasso C, Lee C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics 2004;20: 1546-56.
115. Nuin PAS, Wang Z, Tillier ERM. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 2006;7:471.
116. Roshan U, Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics 2006;22:2715-21.
117. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 2007;23:802-8.
118. Wilm A, Higgins DG, Notredame C. R-Coffee: a method for multiple alignment of non-coding RNA. Nucl. Acids Res 2008; 36:e52.
119. O'Sullivan O, Suhre K, Abergel C, et al. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol 2004;340:385-95.
120. Armougom F, Moretti S, Poirot O, et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucl Acids Res 2006;34:W604-8.
121. Heringa J. Local weighting schemes for protein multiple sequence alignment. Comput Chem 2002;26:459-477.
122. Heringa J. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Comput Chem 1999;23:341-64.
123. Pirovano W, Feenstra KA, Heringa J. PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics 2008;24:492-7.
124. Sahraeian SME, Yoon BJ. PicXAA: greedy probabilistic construction of maximum expected accuracy alignment of multiple sequences. Nucl Acids Res 2010;38:4917-28.
125. Margulies EH, Cooper GM, Asimenos G, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res 2007;17:760-74.