= Multiple sequence alignment =

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences , generally protein , DNA , or RNA . In many cases , the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor . From the resulting MSA , sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences ' shared evolutionary origins . Visual depictions of the alignment as in the image at right illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column , and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment . Multiple sequence alignment is often used to assess sequence conservation of protein domains , tertiary and secondary structures , and even individual amino acids or nucleotides .

Multiple sequence alignment also refers to the process of aligning such a sequence set . Because three or more sequences of biologically relevant length can be difficult and are almost always time @-@ consuming to align by hand , computational algorithms are used to produce and analyze the alignments . MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex . Most multiple sequence alignment programs use heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive .

= = Dynamic programming and computational complexity = =

A direct method for producing an MSA uses the dynamic programming technique to identify the globally optimal alignment solution . For proteins , this method usually involves two sets of parameters : a gap penalty and a substitution matrix assigning scores or probabilities to the alignment of each possible pair of amino acids based on the similarity of the amino acids ' chemical properties and the evolutionary probability of the mutation . For nucleotide sequences a similar gap penalty is used , but a much simpler substitution matrix , wherein only identical matches and mismatches are considered , is typical . The scores in the substitution matrix may be either all positive or a mix of positive and negative in the case of a global alignment , but must be both positive and negative , in the case of a local alignment .

For n individual sequences , the naive method requires constructing the n @-@ dimensional equivalent of the matrix formed in standard pairwise sequence alignment . The search space thus increases exponentially with increasing n and is also strongly dependent on sequence length . Expressed with the big O notation commonly used to measure computational complexity , a naïve MSA takes O (LengthNseqs) time to produce . To find the global optimum for n sequences this way has been shown to be an NP @-@ complete problem . In 1989 , based on Carrillo @-@ Lipman Algorithm , Altschul introduced a practical method that uses pairwise alignments to constrain the n @-@ dimensional search space . In this approach pairwise dynamic programming alignments are performed on each pair of sequences in the query set , and only the space near the n @-@ dimensional intersection of these alignments is searched for the n @-@ way alignment . The MSA program optimizes the sum of all of the pairs of characters at each position in the alignment (the so @-@ called sum of pair score) and has been implemented in a software program for constructing multiple sequence alignments .

= = Progressive alignment construction = =

The most widely used approach to multiple sequence alignments uses a heuristic search known as progressive technique (also known as the hierarchical or tree method) developed by Paulien Hogeweg and Ben Hesper in 1984 . Progressive alignment builds up a final MSA by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related . All progressive alignment methods require two stages : a first stage in which the relationships between the sequences are represented as a tree , called a guide tree , and a second step in which the MSA is built by adding the sequences sequentially to the growing MSA according to the guide tree . The initial guide tree is determined by an efficient clustering method such as neighbor @-@ joining or UPGMA , and may use distances based on the number of identical two letter sub @-@ sequences (as in FASTA rather than a dynamic programming alignment) .

Progressive alignments are not guaranteed to be globally optimal . The primary problem is that when errors are made at any stage in growing the MSA , these errors are then propagated through to the final result . Performance is also particularly bad when all of the sequences in the set are rather distantly related . Most modern progressive methods modify their scoring function with a secondary weighting function that assigns scaling factors to individual members of the query set in a nonlinear fashion based on their phylogenetic distance from their nearest neighbors . This corrects for non @-@ random selection of the sequences given to the alignment program .

Progressive alignment methods are efficient enough to implement on a large scale for many (100s to 1000s) sequences . Progressive alignment services are commonly available on publicly accessible web servers so users need not locally install the applications of interest . The most popular progressive alignment method has been the Clustal family , especially the weighted variant ClustalW to which access is provided by a large number of web portals including GenomeNet , EBI , and EMBNet . Different portals or implementations can vary in user interface and make different parameters accessible to the user . ClustalW is used extensively for phylogenetic tree construction , in spite of the author 's explicit warnings that unedited alignments should not be used in such studies and as input for protein structure prediction by homology modeling . Current version of Clustal family is ClustalW2 . EMBL @-@ EBI announced that CLustalW2 will be expired in August 2015 . They recommend Clustal Omega which performs based on seeded guide trees and HMM profile @-@ profile techniques for protein alignments . They offer different MSA tools for progressive DNA alignments . One of them is MAFFT (Multiple Alignment using Fast Fourier Transform) .

Another common progressive alignment method called T @-@ Coffee is slower than Clustal and its derivatives but generally produces more accurate alignments for distantly related sequence sets . T @-@ Coffee calculates pairwise alignments by combining the direct alignment of the pair with indirect alignments that aligns each sequence of the pair to a third sequence . It uses the output from Clustal as well as another local alignment program LALIGN , which finds multiple regions of local alignment between two sequences . The resulting alignment and phylogenetic tree are used as a guide to produce new and more accurate weighting factors .

Because progressive methods are heuristics that are not guaranteed to converge to a global optimum , alignment quality can be difficult to evaluate and their true biological significance can be obscure . A semi @-@ progressive method that improves alignment quality and does not use a lossy heuristic while still running in polynomial time has been implemented in the program PSAlign .

= = Iterative methods = =

A set of methods to produce MSAs while reducing the errors inherent in progressive methods are classified as " iterative " because they work similarly to progressive methods but repeatedly realign the initial sequences as well as adding new sequences to the growing MSA . One reason progressive methods are so strongly dependent on a high @-@ quality initial alignment is the fact that these alignments are always incorporated into the final result ? that is , once a sequence has been aligned into the MSA , its alignment is not considered further . This approximation improves efficiency at the cost of accuracy . By contrast , iterative methods can return to previously calculated pairwise alignments or sub @-@ MSAs incorporating subsets of the query sequence as a means of optimizing a general objective function such as finding a high @-@ quality alignment score .

A variety of subtly different iteration methods have been implemented and made available in software packages ; reviews and comparisons have been useful but generally refrain from choosing a " best " technique . The software package PRRN / PRRP uses a hill @-@ climbing algorithm to optimize its MSA alignment score and iteratively corrects both alignment weights and locally divergent or " gappy " regions of the growing MSA . PRRP performs best when refining an alignment previously constructed by a faster method .

Another iterative program , DIALIGN , takes an unusual approach of focusing narrowly on local alignments between sub @-@ segments or sequence motifs without introducing a gap penalty . The alignment of individual motifs is then achieved with a matrix representation similar to a dot @-@ matrix plot in a pairwise alignment . An alternative method that uses fast local alignments as anchor points or " seeds " for a slower global @-@ alignment procedure is implemented in the CHAOS / DIALIGN suite .

A third popular iteration @-@ based method called MUSCLE (multiple sequence alignment by log @-@ expectation) improves on progressive methods with a more accurate distance measure to assess the relatedness of two sequences . The distance measure is updated between iteration stages (although , in its original form , MUSCLE contained only 2 @-@ 3 iterations depending on whether refinement was enabled) .

= = Consensus methods = =

Consensus methods attempt to find the optimal multiple sequence alignment given multiple different alignments of the same set of sequences . There are two commonly used consensus methods , M @-@ COFFEE and MergeAlign . M @-@ COFFEE uses multiple sequence alignments generated by seven different methods to generate consensus alignments . MergeAlign is capable of generating consensus alignments from any number of input alignments generated using different models of sequence evolution or different methods of multiple sequence alignment . The default option for MergeAlign is to infer a consensus alignment using alignments generated using 91 different models of protein sequence evolution .

= = Hidden Markov models = =

Hidden Markov models are probabilistic models that can assign likelihoods to all possible combinations of gaps , matches , and mismatches to determine the most likely MSA or set of possible MSAs . HMMs can produce a single highest @-@ scoring output but can also generate a family of possible alignments that can then be evaluated for biological significance . HMMs can produce both global and local alignments . Although HMM @-@ based methods have been developed relatively recently , they offer significant improvements in computational speed , especially for sequences that contain overlapping regions .

Typical HMM @-@ based methods work by representing an MSA as a form of directed acyclic graph known as a partial @-@ order graph , which consists of a series of nodes representing possible entries in the columns of an MSA . In this representation a column that is absolutely conserved (that is , that all the sequences in the MSA share a particular character at a particular position) is coded as a single node with as many outgoing connections as there are possible characters in the next column of the alignment . In the terms of a typical hidden Markov model , the observed states are the individual alignment columns and the " hidden " states represent the presumed ancestral sequence from which the sequences in the query set are hypothesized to have descended . An efficient search variant of the dynamic programming method , known as the Viterbi algorithm , is generally used to successively align the growing MSA to the next sequence in the query set to produce a new MSA . This is distinct from progressive alignment methods because the alignment of prior sequences is updated at each new sequence addition . However , like progressive methods , this technique can be influenced by the order in which the sequences in the query set are integrated into the alignment , especially when the sequences are distantly related .

Several software programs are available in which variants of HMM @-@ based methods have been implemented and which are noted for their scalability and efficiency , although properly using an HMM method is more complex than using more common progressive methods . The simplest is POA (Partial @-@ Order Alignment) ; a similar but more generalized method is implemented in the packages SAM (Sequence Alignment and Modeling System) . and HMMER . SAM has been used as a source of alignments for protein structure prediction to participate in the CASP structure prediction experiment and to develop a database of predicted proteins in the yeast species S. cerevisiae . HHsearch is a software package for the detection of remotely related protein sequences based on the pairwise comparison of HMMs . A server running HHsearch (HHpred) was by far the fastest of the 10 best automatic structure prediction servers in the CASP7 and CASP8 structure prediction competitions .

= = Genetic algorithms and simulated annealing = =

Standard optimization techniques in computer science ? both of which were inspired by , but do not directly reproduce , physical processes ? have also been used in an attempt to more efficiently produce quality MSAs . One such technique , genetic algorithms , has been used for MSA production in an attempt to broadly simulate the hypothesized evolutionary process that gave rise to the divergence in the query set . The method works by breaking a series of possible MSAs into fragments and repeatedly rearranging those fragments with the introduction of gaps at varying positions . A general objective function is optimized during the simulation , most generally the " sum of pairs " maximization function introduced in dynamic programming @-@ based MSA methods . A technique for protein sequences has been implemented in the software program SAGA (Sequence Alignment by Genetic Algorithm) and its equivalent in RNA is called RAGA .

The technique of simulated annealing , by which an existing MSA produced by another method is refined by a series of rearrangements designed to find better regions of alignment space than the one the input alignment already occupies . Like the genetic algorithm method , simulated annealing maximizes an objective function like the sum @-@ of @-@ pairs function . Simulated annealing uses a metaphorical " temperature factor " that determines the rate at which rearrangements proceed and the likelihood of each rearrangement ; typical usage alternates periods of high rearrangement rates with relatively low likelihood (to explore more distant regions of alignment space) with periods of lower rates and higher likelihoods to more thoroughly explore local minima near the newly " colonized " regions . This approach has been implemented in the program MSASA (Multiple Sequence Alignment by Simulated Annealing) .

= = Phylogeny @-@ aware methods = =

Most multiple sequence alignment methods try to minimize the number of insertions / deletions (gaps) and , as a consequence , produce compact alignments . This causes several problems if the sequences to be aligned contain non @-@ homologous regions , if gaps are informative in a phylogeny analysis . These problems are common in newly produced sequences that are poorly annotated and may contain frame @-@ shifts , wrong domains or non @-@ homologous spliced exons .

The first such method was developed in 2005 by Löytynoja and Goldman . The same authors released a software package called PRANK in 2008 . PRANK improves alignments when insertions are present . Nevertheless , it runs slowly compared to progressive and / or iterative methods which have been developed for several years .

In 2012 , two new phylogeny @-@ aware tools appeared . One is called PAGAN that was developed by the same team as PRANK . The other is ProGraphMSA developed by Szalkowski . Both software packages were developed independently but share common features , notably the use of graph algorithms to improve the recognition of non @-@ homologous regions , and an improvement in code making these software faster than PRANK .

= = Motif finding = =

Motif finding , also known as profile analysis , is a method of locating sequence motifs in global MSAs that is both a means of producing a better MSA and a means of producing a scoring matrix for use in searching other sequences for similar motifs . A variety of methods for isolating the motifs have been developed , but all are based on identifying short highly conserved patterns within the larger alignment and constructing a matrix similar to a substitution matrix that reflects the amino acid or nucleotide composition of each position in the putative motif . The alignment can then be refined using these matrices . In standard profile analysis , the matrix includes entries for each possible character as well as entries for gaps . Alternatively , statistical pattern @-@ finding algorithms can identify motifs as a precursor to an MSA rather than as a derivation . In many cases when the query set contains only a small number of sequences or contains only highly related sequences , pseudocounts are added to normalize the distribution reflected in the scoring matrix . In particular , this corrects zero @-@ probability entries in the matrix to values that are small but nonzero .

Blocks analysis is a method of motif finding that restricts motifs to ungapped regions in the alignment . Blocks can be generated from an MSA or they can be extracted from unaligned sequences using a precalculated set of common motifs previously generated from known gene families . Block scoring generally relies on the spacing of high @-@ frequency characters rather than on the calculation of an explicit substitution matrix . The BLOCKS server provides an interactive method to locate such motifs in unaligned sequences .

Statistical pattern @-@ matching has been implemented using both the expectation @-@ maximization algorithm and the Gibbs sampler . One of the most common motif @-@ finding tools , known as MEME , uses expectation maximization and hidden Markov methods to generate motifs that are then used as search tools by its companion MAST in the combined suite MEME / MAST .

= = Non @-@ Coding Multiple Sequence Alignment = =

Non @-@ coding DNA regions , especially TFBSs , are rather more conserved and not necessarily evolutionarily related , and may have converged from non @-@ common ancestors . Thus , the assumptions used to align protein sequences and DNA coding regions are inherently different from those that hold for TFBS sequences . Although it is meaningful to align DNA coding regions for homologous sequences using mutation operators , alignment of binding site sequences for the same transcription factor cannot rely on evolutionary related mutation operations . Similarly , the evolutionary operator of point mutations can be used to define an edit distance for coding sequences , but this has little meaning for TFBS sequences because any sequence variation has to maintain a certain level of specificity for the binding site to function . This becomes specifically important when trying to align known TFBS sequences to build supervised models to predict unknown locations of the same TFBS . Hence , Multiple Sequence Alignment methods need to adjust the underlying evolutionary hypothesis and the operators used as in the work published incorporating neighbouring base thermodynamic information to align the binding sites searching for the lowest thermodynamic alignment conserving specificity of the binding site , EDNA .

= = Alignment visualization and quality control = =

The necessary use of heuristics for multiple alignment means that for an arbitrary set of proteins , there is always a good chance that an alignment will contain errors . For example , an evaluation of several leading alignment programs using the BAliBase benchmark found that at least 24 % of all pairs of aligned amino acids were incorrectly aligned . These errors can arise because of unique insertions into one or more regions of sequences , or through some more complex evolutionary process leading to proteins that do not align easily by sequence alone . As the number of sequence and their divergence increases many more errors will be made simply because of the heuristic nature of MSA algorithms . Multiple sequence alignment viewers enable alignments to be visually reviewed , often by inspecting the quality of alignment for annotated functional sites on two or more sequences . Many also enable the alignment to be edited to correct these (usually minor) errors , in order to obtain an optimal ' curated ' alignment suitable for use in phylogenetic analysis or comparative modeling .

However , as the number of sequences increases and especially in genome @-@ wide studies that involve many MSAs it is impossible to manually curate all alignments . Furthermore , manual curation is subjective . And finally , even the best expert cannot confidently align the more ambiguous cases of highly diverged sequences . In such cases it is common practice to use automatic procedures to exclude unreliably aligned regions from the MSA . For the purpose of phylogeny reconstruction (see below) the Gblocks program is widely used to remove alignment blocks suspect of low quality , according to various cutoffs on the number of gapped sequences in alignment columns . However , these criteria may excessively filter out regions with insertion / deletion events that may still be aligned reliably , and these regions might be desirable for other purposes such as detection of positive selection . A few alignment algorithms output site @-@ specific scores that allow the selection of high @-@ confidence regions . Such a service was first offered by the SOAP program , which tests the robustness of each column to perturbation in the parameters of the popular alignment program CLUSTALW . The T @-@ Coffee program uses a library of alignments in the construction of the final MSA , and its output MSA is colored according to confidence scores that reflect the agreement between different alignments in the library regarding each aligned residue . Its extension , TCS : (Transitive Consistency Score) , uses T @-@ Coffee libraries of pairwise alignments to evaluate any third party MSA . Pairwise projections can be produced using fast or slow methods , thus allowing a trade @-@ off between speed and accuracy . Another alignment program that can output an MSA with confidence scores is FSA , which uses a statistical model that allows calculation of the uncertainty in the alignment . The HoT (Heads @-@ Or @-@ Tails) score can be used as a measure of site @-@ specific alignment uncertainty due to the existence of multiple co @-@ optimal solutions . The GUIDANCE program calculates a similar site @-@ specific confidence measure based on the robustness of the alignment to uncertainty in the guide tree that is used in progressive alignment programs . An alternative , more statistically justified approach to assess alignment uncertainty is the use of probabilistic evolutionary models for joint estimation of phylogeny and alignment . A Bayesian approach allows calculation of posterior probabilities of estimated phylogeny and alignment , which is a measure of the confidence in these estimates . In this case , a posterior probability can be calculated for each site in the alignment . Such an approach was implemented in the program BAli @-@ Phy .

There free available programs for visualization of multiple sequence alignments : JalView , UGENE .

= = Use in phylogenetics = =

Multiple sequence alignments can be used to create a phylogenetic tree . This is made possible by two reasons . The first is because functional domains that are known in annotated sequences can be used for alignment in non @-@ annotated sequences . The other is that conserved regions known to be functionally important can be found . This makes it possible for multiple sequence alignments to be used to analyze and find evolutionary relationships through homology between sequences . Point mutations and insertion or deletion events (called indels) can be detected .

Multiple sequence alignments can also be used to identify functionally important sites , such as binding sites , active sites , or sites corresponding to other key functions , by locating conserved domains . When looking at multiple sequence alignments , it is useful to consider different aspects of the sequences when comparing sequences . These aspects include identity , similarity , and homology . Identity means that the sequences have identical residues at their respective positions . On the other hand , similarity has to do with the sequences being compared having similar residues quantitatively . For example , in terms of nucleotide sequences , pyrimidines are considered similar to each other , as are purines . Similarity ultimately leads to homology , in that the more similar sequences are , the closer they are to being homologous . This similarity in sequences can then go on to help find common ancestry .

= = = Survey articles = = =

Duret , L. ; S. Abdeddaim (2000) . " Multiple alignment for structural functional or phylogenetic analyses of homologous sequences " . In D. Higgins and W. Taylor . Bioinformatics sequence structure and databanks . Oxford : Oxford University Press .

Notredame , C. (2002) . " Recent progresses in multiple sequence alignment : a survey " . Pharmacogenomics 31 (1) : 131 ? 144 @.@ doi : 10 @.@ 1517 / 14622416 @.@ 3 @.@ 1 @.@ 131 . PMID 11966409 .
Thompson , J. D. ; Plewniak , F. ; Poch , O. (1999) . " A comprehensive comparison of multiple sequence alignment programs " . Nucleic Acids Research 27 (13) : 12682 ? 2690 @.@ doi : 10 @.@ 1093 / nar / 27 @.@ 13 @.@ 2682 . PMC 148477 . PMID 10373585 .
Wallace , I.M. ; Blackshields , G. ; Higgins , D.G. (2005) . " Multiple sequence alignments " . Curr Opin Struct Biol 15 (3) : 261 ? 266 @.@ doi : 10 @.@ 1016 / j.sbi.2005.04.002. PMID 15963889 .
Notredame , C (2007) . " Recent Evolutions of Multiple Sequence Alignment Algorithms " . PLOS Computational Biology 8 (3) : e123. doi : 10 @.@ 1371 / journal.pcbi.0030123. PMC 1963500 . PMID 17784778 .

= = = Lecture notes , tutorials , and courses = = =

Multiple sequence alignment lectures ? from the Max Planck Institute for Molecular Genetics
Lecture Notes and practical exercises on multiple sequence alignments at the EMBL
Molecular Bioinformatics Lecture Notes
Molecular Evolution and Bioinformatics Lecture Notes