Multiple Alignment

Julie Thompson, *Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France*

Olivier Poch, *Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg, France*

Multiple alignment is a powerful integrative tool that addresses a variety of biological problems, ranging from key functional residue detection to the evolution of a protein family. Traditionally, a multiple alignment was generally constructed as a series of pairwise alignments; however, the recent application of various new computational techniques to the multiple alignment problem has led to a number of interesting new developments.

Advanced article

Article contents

doi: 10.1038/npg.els.0005258

Introduction

Since its introduction in the early 1970s, multiple-sequence alignment has become a fundamental tool in a number of different domains in modern molecular biology. Multiple alignments present a synthetic view of the variability along the sequence and among families of homologous sequences, thus providing a reliable context in which to include and compare distant homologs. Evolutionary studies based on sequence data rely on multiple alignments to define the phylogenetic relationships between organisms. Multiple alignments are also invaluable for homology structure modeling. Sequence similarity between proteins usually indicates a structural resemblance, and accurate sequence alignments provide a practical approach for both two-dimensional (2D) and three-dimensional (3D) structure modeling. The multiple alignment also highlights conserved structural motifs or key functional residues that characterize a family of proteins. This is crucial for experimental biologists in the determination of catalytic residues or residues involved in interactions in a new family of proteins. It is also vital in drug design to specifically target an essential protein on the basis of biochemical properties unique to a precise pathogen group. Traditionally, most multiple-alignment programs were based on dynamic programming algorithms, similar to those used for pairwise sequence alignments. However, the multiple alignment of the highly complex proteins detected by today's advanced database search methods is a daunting task, and there has been renewed interest in the application of novel computational techniques to solve the multiple-alignment problem. The most recent approaches have moved away from a single, all-encompassing algorithm to a more cooperative strategy, integrating different, complementary algorithms and/or incorporating biological information other than the sequence itself. (***See*** Gene Feature Identification; Protein Homology Modeling; Sequence Similarity.)

Progressive Multiple Alignment

The first formal algorithm for multiple-sequence alignment (Sankoff, 1975) extended the basic pairwise dynamic programming algorithm to multiple sequences. However, the optimal, exact alignment of more than a few sequences (more than 10) remained impractical due to the intensive computer resources required, despite some recent space and time improvements. Heuristic approaches were required to reduce the problem to a reasonable size. One of the first heuristic methods, which is still in widespread use today, exploits the fact that homologous sequences are evolutionarily related. A multiple alignment is built up progressively by a series of pairwise alignments, following the branching order in a phylogenetic tree (Feng and Doolittle, 1987). An example using five immunoglobulin-like domains is shown in **Figure 1**. (***See*** Dynamic Programming; Sequence Alignment.)

This procedure works well when the sequences to be aligned are of different degrees of divergence. Pairwise alignment of closely related sequences can be performed very accurately. By the time the more distantly related sequences are aligned, important information about the variability at each position is available from those sequences already aligned. A number of different alignment programs based on this method exist, using either a global alignment method to construct an alignment of the complete sequences or a local algorithm to align only the most conserved subsegments of the sequences. They differ mainly in the method used to determine the order of alignment of the sequences (**Table 1**). Since then, the sensitivity of the progressive multiple-sequence alignment method has been somewhat improved, with the introduction of several important enhancements to the basic method. For example, Treealign (Hein, 1990) extends the progressive alignment process by adding a parsimony step: an initial alignment is constructed and used to build a parsimony tree, which in turn is used to direct

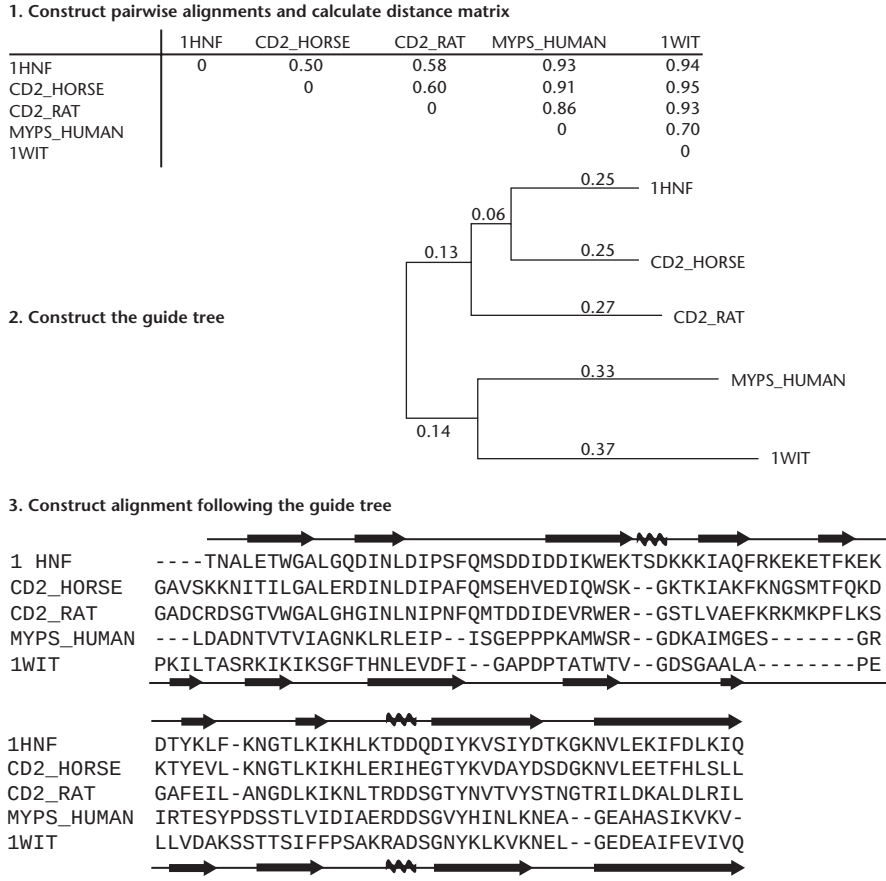

Figure 1 The basic progressive alignment procedure, exemplified by a set of five immunoglobulin-like domains. The sequence names are from the SWISS-PROT or Protein Data Bank (PDB) databases: 1HNF, human cell adhesion (CD2) protein; CD2_HORSE, horse cell adhesion protein; CD2_RAT, rat cell adhesion protein; MYPS_HUMAN, human myosin-binding protein; 1WIT, nematode twitchin muscle protein. The first step involves aligning all possible pairs of sequences in order to determine the distances between them. A guide tree is then created and is used to determine the order of the multiple alignment. First, the human and horse CD2 sequences are aligned. These two sequences are then aligned with the rat CD2 sequence. Finally, the myosin-binding protein sequence is aligned with the twitchin sequence, before being merged with the alignment of the three CD2 sequences. The secondary structure elements of the immunoglobulin-like domains from the human CD2 (1HNF) and the nematode twitchin (1WIT) proteins are shown above and below the alignment (right arrow, beta sheet; coil, alpha helix).

the final alignment algorithm. CLUSTAL_X (Thompson *et al.*, 1997) reduces the problem of the overrepresentation of certain sequences by incorporating a sequence-weighting scheme that downweights near-duplicate sequences and upweights the most divergent ones. In addition, position-specific gap penalties encourage the alignment of new gaps on existing gaps introduced earlier in the multiple alignment. Most of the alignment programs mentioned here use one residue scoring matrix and two gap penalties (one for opening a new gap and one for extending an existing gap). When identities dominate an alignment, almost any set of parameters will find approximately the correct solution. With very divergent sequences, however, the scores given to nonidentical residues will become critically important. Also, the exact values of the gap penalties become important for success. Thus, the choice of alignment parameters remains a decisive factor affecting the quality of the final alignment. (***See*** Global Alignment; Substitution Matrices.)

Iterative Strategies

While the above methods, based on dynamic programming, have proved relatively successful in providing accurate multiple alignments of sequences that are related over their entire lengths or contain relatively well-conserved regions, the multiple-alignment problem is becoming more complex. Global alignment

Table 1 Some of the most widely used multiple-sequence alignment programs

Program	Authors	Algorithm	Local/global	Alignment ordering[a]
Msa	Gupta SK, Kececioglu JD and Schaffer AA (1995)	Optimal dynamic programming	Global	N/A
Dca	Stoye J (1998)	Optimal dynamic programming	Global	N/A
Pima	Smith RF and Smith TF (1992)	Progressive	Local	SB, ML
Multalign	Barton GJ and Sternberg JE (1987)	Progressive, iterative refinement	Global	SB
Pileup	Wisconsin Package, Genetics Computer Group, Madison, WI, USA	Progressive	Global	UPGMA
CLUSTAL_X	Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG (1997)	Progressive	Global	NJ
Prrp	Gotoh O (1996)	Iterative refinement	Global	N/A
Sam-T98	Karplus K, Barrett C and Hughey R (1998)	HMM	Local	N/A
Hmmer	Eddy SR (1998)	HMM	Global	N/A
Saga/Raga	Notredame C and Higgins DG (1996)	Genetic algorithm	Global	N/A
MACAW	Schuler GD, Altschul SF and Lipman DJ (1991)		Local	N/A
Probe	Neuwald AF, Liu JS, Lipman DJ and Lawrence CE (1997)	Simulated annealing	Local	N/A
Dialign	Morgenstern B (1999)	Segment-to-segment	Local	N/A

[a]The method used to determine the order of the progressive multiple alignment: SB, sequential branching; ML, maximum likelihood; UPGMA, unweighted pair grouping method; NJ, neighbor joining; N/A, not applicable.

of multidomain proteins, often containing large *N*- and/or *C*-terminal extensions and/or internal insertions is becoming a standard requirement. This has aroused new interest in the multiple-alignment problem, and a number of interesting new developments have recently been reported (**Table 1**). A common point of interest has been the application of iterative strategies to refine and improve the initial alignment. The protein sequence information–basic local alignment search tool (PSI-BLAST) program builds multiple alignments by aligning the homologs detected by a BLAST database search to the query sequence. Hidden Markov models (HMMs) have been used in a number of programs to build multiple alignments and have been employed notably to create large reference databases of sequence alignments such as Pfam and PROSITE. The flexibility and efficiency of stochastic techniques such as simulated annealing and genetic algorithms (SAGA) have also been exploited in the search for more accurate alignments. Iteration techniques have also been used to refine an initial multiple alignment built using the traditional progressive alignment algorithm. An alternative to the global alignment approach is the 'segment-to-segment' alignment method. Segments consisting of locally conserved residue patterns or motifs, rather than individual residues, are detected and then combined to construct a local multiple alignment of only the most conserved regions of the sequences. (***See*** BLAST Algorithm; Hidden Markov Models; Sequence Complexity and Composition; Similarity Search.)

Cooperative Strategies

The complexity of the multiple-alignment problem has led to the combination of different alignment algorithms and the incorporation of biological information other than the sequence itself. ComAlign (Bucka-Lassen *et al*., 1999) extracts qualitatively good subalignments from a set of multiple alignments and combines these into a new, often improved alignment. T-Coffee (Notredame *et al*., 2000) incorporates information from heterogeneous data sources such as local and global alignments, structure alignments or known motifs in a progressive multiple alignment. In the case of the DbClustal program (Thompson *et al*., 2000), locally conserved segments are mined from the sequence databases and are then used to guide the global multiple alignment. Methods have also been developed that combine primary sequence and 2D or 3D structure information to produce a single multiple alignment, for example Heringa (1999). Thus, information other than the

sequences themselves is now being incorporated into the multiple alignment in an effort to improve alignment accuracy. (***See*** Protein Structure.)

Assessing Multiple Alignment Quality

Although significant improvement in alignment quality has been reported for many of these new programs, the lack of a standard benchmarking system has hindered an objective evaluation of the diverse algorithms. However, some progress is now being made in this area. Starting in 1994, 12 different multiple-alignment programs were compared (McClure *et al*., 1994), using an alignment benchmark consisting of four different sets of protein sequences, and it was concluded that global alignment algorithms generally performed better than local methods. Then, with the growth of the protein structure databases, it became standard practice to compare the results of multiple-alignment programs to 'standard-of-truth' alignments based on 3D structural superpositions. A recent study (Thompson *et al*., 1999) of many of the programs mentioned above, using a benchmark alignment database that was specifically designed for the evaluation of multiple-alignment programs, identified a number of characteristic features of the various algorithms. The comparison showed that, while global alignment methods in general performed better for sets of sequences that were of similar length, local algorithms were more successful at identifying the most conserved motifs in sequences containing large extensions and insertions. The same study also showed that the new iterative algorithms often produced more accurate alignments, although at the expense of a heavy time penalty. (***See*** Protein Structure Prediction and Databases.)

Objective functions

In the absence of an accurate reference alignment, such as those based on 3D structures, it is still necessary to estimate the quality or reliability of an alignment. Most multiple-alignment methods define a scoring function that assigns a numerical value to each possible alignment and attempts to maximize this score. However, the 'optimal' alignment defined by a multiple-alignment method is not necessarily the same as the 'biologically correct' alignment. The correctness of an alignment has often been evaluated manually by an expert, taking into account the conservation of motifs or secondary structure elements. But, for high-throughput biology such as genome annotation and analysis projects, a reliable and automatic scoring method that accurately reflects the biological quality of an alignment is essential. One of the first, and most popular, scoring schemes for multiple alignments was the sum-of-pairs (SP) score (Carrillo and Lipman, 1988), where the score for a multiple alignment is simply the sum of the scores for all pairwise alignments. A number of variations on the original SP score exist, including the use of sequence weights and various gap penalty schemes. More recent work has concentrated on column statistics, for example minimum entropy, maximum likelihood scores and the mean distance scores introduced in ClustalX. These measures, also known as objective functions, are currently used to evaluate and compare multiple alignments from different sources. They are also used in iterative alignment methods to improve the alignment by seeking to maximize the objective function. However, the search for a reliable function that genuinely reflects the biological significance of an alignment could be compared to the search for the Holy Grail. (***See*** Alignment: Statistical Significance.)

Perspectives

A more recent application of sequence alignments has been in genome annotation projects. As the number of completely sequenced genomes rapidly increases, the number of proteins in the sequence databases with no functional or structural annotation is becoming a serious problem. Sensitive methods of sequence analysis are crucial in order to extract as much functional information as possible from the genomic sequence data. The classic approach consists of searching the databases to derive functional and structural information from previously annotated homologs. Global multiple alignment of the detected homologs constitutes a complementary step in functional assignment where quality control can take place. The application of multiple alignments at the genome level also opens the way to the phylogenetic analysis of complete proteomes and to the study of the coevolution of sets of proteins. Further, global multiple alignments permit more detailed sequence analysis, such as verification of the reading frame lengths and determination of the domain organization of a protein family. Unfortunately, multiple alignments have often been considered unsuitable for high-throughput analysis of genomic sequences because of their unreliability in the face of complex, often noncollinear proteins. Despite the recent advances resulting from the new multiple-alignment techniques, a number of problems remain to be solved. Large multidomain proteins are becoming more and more prevalent, in particular with the arrival of a number of genome sequences from eukaryotic organisms. Proteins with nonlinear elements such as repeats, inversions and circular permutations, or low-complexity regions such as

transmembrane proteins or coiled coils, cause particular problems for multiple-alignment programs. Clearly, an accurate multiple alignment can no longer be constructed from the primary sequence data alone. No single algorithm currently exists that can cope with the highly complex proteins detected by today's database search programs. The way forward is undoubtedly an integrated system that will bring together knowledge-based or text-mining systems and prediction methods, with their inherent unreliability. The incorporation of heterogeneous, often inconsistent data will require major changes to the fundamental alignment algorithms used to date. Now that public access to the wealth of biological data is possible due to the widespread adoption of the internet network as a standard research tool, such integration has become a realistic objective. However, all this must still be achieved within the timescale exacted by the high-throughput genome projects. (***See*** Multidomain Proteins; Protein Databases; Protein Families: Evolution; Protein Sequence Databases.)

See also

Alignment: Statistical Significance
Global Alignment
Sequence Alignment

References

Barton GJ and Sternberg MJ (1987) A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *Journal of Molecular Biology* **198**: 327–337.

Bucka-Lassen K, Caprani O and Hein J (1999) Combining many multiple alignments in one improved alignment. *Bioinformatics* **15**: 122–130.

Carrillo H and Lipman D (1988) The multiple sequence alignment problem in biology. *SIAM Journal of Applied Mathematics* **48**: 1073–1082.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Feng DF and Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**: 351–360.

Gotoh O (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *Journal of Molecular Biology* **264**: 823–838.

Gupta SK, Kececioglu JD and Schaffer AA (1995) Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *Journal of Computational Biology* **2**: 459–472.

Hein J (1990) Unified approach to alignment and phylogenies. *Methods in Enzymology* **183**: 626–645.

Heringa J (1999) Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Computational Chemistry* **23**: 341–364.

Karplus K, Barrett C and Hughey R (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**: 846–856.

McClure MA, Vasi TK and Fitch WM (1994) Comparative analysis of multiple protein sequence alignment methods. *Molecular Biology and Evolution* **11**: 571–592.

Morgenstern B (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**: 211–218.

Neuwald AF, Liu JS, Lipman DJ and Lawrence CE (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Research* **25**: 1665–1677.

Notredame C and Higgins DG (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Research* **24**: 1515–1524.

Notredame C, Higgins DG and Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**: 205–217.

Sankoff D (1975) Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics* **78**: 35–42.

Schuler GD, Altschul SF and Lipman DJ (1991) A workbench for multiple alignment construction and analysis. *Proteins* **9**: 180–190.

Smith RF and Smith TF (1992) Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Engineering* **5**: 35–41.

Stoye J (1998) Multiple sequence alignment with the divide-and-conquer method. *Gene* **211**: GC45–GC56.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**: 4876–4882.

Thompson JD, Plewniak F and Poch O (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research* **27**: 2682–2690.

Thompson JD, Plewniak F, Thierry JC and Poch O (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Research* **28**: 2919–2926.

Further Reading

Baxevanis AD (1998) Practical aspects of multiple sequence alignment. *Methods in Biochemical Analysis* **39**: 172–188.

Durbin R, Eddy S, Krogh A and Mitchison G (1999) Multiple sequence alignment methods. In: Durbin R (ed.) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, pp 134–159. Cambridge, UK: Cambridge University Press.

Duret L and Abdeddaim S (2000) Multiple alignments for structural, functional, or phylogenetic analysis of homologous sequences. In: Higgins DG and Taylor WR (eds.) *Bioinformatics: Sequence, Structure, and Databanks: A Practical Approach*, pp 51–76. Oxford, UK: Oxford University Press.

Gonnet GH, Korostensky C and Benner S (2000) Evaluation measures of multiple sequence alignments. *Journal of Computational Biology* **7**: 261–276.

Gotoh O (1999) Multiple sequence alignment: algorithms and applications. *Advanced Biophysics* **36**: 159–206.

Higgins DG and Taylor WR (2000) Multiple sequence alignment. *Methods in Molecular Biology* **143**: 1–18.

Hirosawa M, Totoki Y, Hoshida M and Ishikawa M (1995) Comprehensive study on iterative algorithms of multiple sequence alignment. *Computer Applications in the Biosciences* **11**: 13–18.

Phillips A, Janies D and Wheeler W (2000) Multiple sequence alignment in phylogenetic analysis. *Molecular and Phylogenetic Evolution* **16**: 317–330.