Volume 6, Issue 3, March 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Literature Survey on Multiple Sequence Alignment Algorithms

[1]Lakshmi Naga Jayaprada. Gavarraju, [2]P. Jeevana Jyothi, [3]K. Karteeka Pawan

[1] Department of CSE, Narasaraopeta Engineering College, Narasaraopet, 522 601, Andhra Pradesh, India

[2] Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, 522 509, Andhra Pradesh, India\

[2] Department of IT, Guntur, 522 019, Andhra Pradesh, India

Abstract –Multiple sequence alignment is the task of recognizing evolutionarily or structurally related positions in a collection of multiple amino acid sequences.That means it can tell us about the evolution of the organisms, it can see which regions of a gene are prone to mutation and which can have one residue substituted by another without altering function, it is possible to study Homologous genes and can discover paralogs and orthologs genes that are evolutionary related.Though the multiple sequence alignmentproblem has been studied for several decades, many modern studies have demonstrated significantgrowth in refining the accuracy or scalability of multiple and pairwise sequence alignment algorithms. In this paper, the review of state-of-the-art multiple sequence alignment algorithms which were developed up to now are specified.

Key Words: Sequence Analysis; Multiple sequence alignment; Dynamic Programming; Local Alignment; Global Alignment.

I. INTRODUCTION

Multiple sequence alignment (MSA)[1] is the dominantmethod for deducing biological facts from a set of sequences. It encompasses the alignment of more than two sequences and aims to discoveralike positions across the aligned query set of sequences. An MSA can be observedas a 2-dimensional table in which the sequences are the rows and the columns of equivalent amino acids have been arranged by placing gap characters in suitable positions, such that the biological relationship of the sequences is best characterized. An MSA can provide a treasure of information about structure-function relationships inside a set of protein sequences, e.g., the evolutionary conservation of functionally or structurally important amino acids at certain sequence positions or conserved hydrophobicity patterns in precise regions. MSAs are an essential requirement to many computational approaches of investigation of protein families, such as homology modelling, secondary structure prediction, and phylogenetic reconstruction. An MSA is achallenge to signify evolutionarily associated sequences in the most reliableway. Even with the considerable history of MSA the methodology to find good sequence alignment is still under continuous development. It is because of the complex relationship often exists among homologous sequences, combined with a lack of information about their accurate evolutionary history, absolute certainty about the perfection of MSAs is often hard to achieve.

With the completion of the first draft of thehuman genome and well over one hundred genomes of other species, the precise alignment of biological sequences has become more important than ever. This is due to the fact that the direct prediction of a protein's structure and function is still a major unresolved problem. Toincrease the familiarity of the function and interaction of protein sequences obtained by sequencing techniques, many initiatives are underway for large-scale proteomics and structure elucidation of novel genomic proteins. However, roughly 50% of the proteins in most sequenced species do not have an assigned function, and consequently an important target of bioinformatics method development is aimed at gathering the function of an increased fraction of translated proteins by enhancing comparative sequence techniques. In the pursuit for knowledge about the role of a certain unknown protein in the cellular molecular network, comparing the query sequence with the many sequences in annotated protein sequence databases often leads to useful suggestions regarding the protein's 3-dimensional (3-D) structure or molecular function. Significantgrowth has been made in homology searching over the last few years by usingMSA techniques in iterative sequence database search strategies [2]. Many current research projects aim to improve the sensitivity of (multiple) sequence alignment techniques, which require high-performance computing given the current and briskly growing database sizes. Given the plethora of sequence data, thealignment engines also have to be tremendouslyfast and fully automatic to be included in genomic pipelines.

Before boarding on a review of the large number of approaches to multiple sequence alignment, it is important to specify that each ofthe methods comes with individual strengths and weaknesses relating to accuracy, sensitivity, speed, consistency, versatility, and the like. There is no single best method for each individual query set of sequences. The user should therefore trya number of approaches, preferably including different strategies, to approach the biological complexity of most MSA problems.

II. GLOBAL OR LOCAL METHODS

Many MSA techniques accomplish*global*alignment [3] andmatch sequences over their complete lengths. Difficulties with this approach can arise when sequencesthat are only homologous over localregions are matched. In such situations, globalalignment techniques might be unsuccessful to identifyextremely similar internal regions because thesemay be dominated by divergent stretchesand high gap penalties normally required toachieve proper global matching. Moreover,many biological sequences are modular andshow shuffled domains [4], which can extract a global alignment oftwo complete sequences meaningless. The occurrenceof varying numbers of internal sequencerepeats [5] can also strictlylimit the applicability of global methods.Generally, when there is a large difference inthe lengths of two sequences to be compared,it is desirable to take account of local alignment techniquesin the analysis. To address these problems,Smith and Waterman [6] developed a so-called *local* alignment techniquein which the most similar regions in twosequences are carefully chosen and aligned. For multiple sequences, the main automatic methods include the Gibbs sampler [7], MEME [8] and Dialign2 [9]. These local MSA programs often perform well when there is a clear block of ungapped alignment commonto all of the sequences, but perform poorly under moderate gap requirements and show inferior results over general sets of test cases when comparedwith global methods [10][11].

III. ACCOMPLISHMENT OF MSA

Performing an MSA on a given set of protein sequence and extracting maximum information from the alignment comprises a number of prominent steps:

1. The selection of sequences.
2. The choice of the scoring function used to compare sequences or sequence blocks.
3. The application and optimization of this scoring function in compiling the alignment.

Selecting the Sequences

An MSA can be ambiguous when a sequence set contains sequences that are not homologous. Ideally, the sequences should all be orthologs, but in practice it is frequently difficult to ensure that this is the case. It should be stressed that most MSA routines will produce an alignment even in the case of biologically unrelated sequences, which can give rise to falserecommendations regarding the proteins' structure or function. A commonly used way to create a sequence set around a given query sequence of interest is to make use of a homology searching technique [12] to scour sequences in public sequence databases.

The Scoring Function

The scoring function is the formalization of the biological knowledge used in aligning the sequences. Ideally, it should contain all available knowledge about evolutionary, structural, and functional aspects of the compared sequences, so that the scoring function estimatesthe biological reality. In practice, however, this information is often not available or cannot be formalized mathematically. Although each cross-comparison of a residue between two sequences should in reality be calculatedindividually based on its structural andfunctional context, the most widely usedscheme to compare sequences is based on generalizedaverages for scoring each pair of residuetypes, given in the form of a symmetric20×20 amino acid exchange matrix. Thescheme models the alignment of two sequencesas a Markov process, where the amino acidmatches are considered independent, so that theproduct of the probabilities for each matchwithin an alignment can be taken.

Applying the Scoring Function

Apart from being a fundamental geneticchallenge, MSA is also a computationally intense problem, which means that for all but the smallest data sets of less than 10 sequences, an exact solution is not realistic. Algorithms that perform simultaneous alignment over a multidimensional search matrix, where each sequence in the MSA represents an extra dimension [13][14], come closest to an exact solution.The most populous class of algorithms is that of progressive MSA methods. The progressive strategy infers that an algorithm for pairwisesequence alignment is recurrently used in a stepwise fashion until all sequences are aligned. In the huge majority of progressive methods the Dynamic Programming (DP) strategy is approved. The DP strategy guarantees that, given an amino acid exchange matrix and gap penalty values, the highest scoring or optimal pairwise alignment is calculated. The progressive alignment strategy reuses the pairwise DP algorithm in a "greedy" manner; i.e., alignments formed during the progression towards the final MSA cannot be changed anymore. The main difference between the available DPbasedmethods is the way in which the information of aligned blocks of sequences is represented. While early methods used consensus sequences to represent alignment blocks, current methods all use a profile formalism to represent the information in an MSA [15]. Recent developments in multiple alignment techniques have mainly concentrated on sensitive and optimal models to represent MSA information. A class of practises that are able to revisitand optimize is that of iterative multiple alignment techniques. Iterative techniques [16] attempt toenhance the alignment quality by gleaninginformation from a multiple alignment assembledin an earlier round, which is then useful in a next round to improve the alignment according to a given scoring scheme.Another classes of alignments is stochastic alignments, where probabilistic frameworks such as hidden Markov models and Bayesian networks have been attempted. Other techniques based on fast computationaltechniques such as suffix trees and fast Fourier transforms (FFT).

IV. DIFFERENT METHODS OF MSA

MSA is a complex problem and over the past 4 decades an increasing number of methods have been established that try to solveit, each with their own strengths and weaknesses.A summary of the properties and accessibility of the various methods is discussed as follows.

BioPat

BioPat is a mathematical package that comprises the first-ever integrated MSA method, which is a global progressive algorithm with iteration abilities [22]. Initially, a coalescence tree (dendrogram) is constructed based on all pairwise similarities of the sequences to be aligned, matchedby dynamic programming.The method provides choices among many of the commonly used clusteringtechniques to build the dendrogram, such asUnweighted Pair-Group Mean Average (UPGMA), the presentday ancestor method orthe Neighbor-Joining (NJ) method. Once the dendrogram is completed, the sequences are progressively aligned following the branch order of the dendrogram. The resulting alignment is then used to infer the associated new pairwise similarities and the initial dendrogram is updated to produce a new alignment. A new dendrogram is then constructed iteratively, from which a succeeding alignment is created based on the increased information.

MultAlin

The method MultAlin [23] follows the Hogeweg and Hesper stylein that it uses hierarchical clustering for building a guide dendrogram and iteration. For the alignment of two sets of sequences it uses the average similarity score between a pair of alignment columns, one from each set, which is the average over the amino acid exchange values connectedwith all pairwise intercolumn residue comparisons. This way of scoring alignment positions effectively follows the profile comparison technique.

MULTAL

The early method MULTAL [24] is very fast and constructs a dendrogram during the progressive alignment, as in the method of Feng and Doolittle. It uses a fast sequential branching method to align the closest pairs of sequences first and then subsequently align the next closest sequences to those already aligned. The order in which the sequences are aligned is largely based on the global amino acid composition of the sequences, which saves the fixed cost of performing all-against-all pairwise alignments. Blocks of aligned sequences are scored by dynamic programming similar to the method MultAlin, but the similarity of two alignment columns is additionally normalized by the minimum number of sequences in either of two compared alignment blocks.

MultAlign

The global progressive method MultAlign [25] establishes a simple chain order in which the individual sequences are aligned one by one. Initially, all pairwise alignment scores are determined and the two most similar sequences are matched first. During further iterations, the sequence showing the highest alignment score when matched with the prealigned sequence block is added to it.The MultAlign method incorporates iteration capabilities in that the resulting MSA can be progressively refined by realigning each sequence with the previous alignment from which that sequence is deleted i.e., sequence A1 is matched with aligned sequences A2...A*N*; sequence A2 is then realigned with the alignment of A1, A3...A*N*, and so forth. This process is repeated until all *N* sequences are realigned.

ClustalW, ClustalX

ClustalW and the later window graphic user interface (GUI) version ClustalX are the newest versions of the global progressive alignment algorithm Clustal [26], and are generally considered as the standard method for MSA. The progressive strategy used is a simplification of the original Feng and Doolittle scheme.The alignment is constructed by first building a guide dendrogram using Neighbor-Joining, based on sequence similarity, which is subsequently used to order successive pairwise alignments. The already aligned sequences are reduced to a profile for the subsequent pairwise alignment. However, during the progressive alignment process, highly specialized heuristics are applied to try and optimize how the sequence information is processed. When the sequences are ordered for alignment according to the precomputed dendrogram, the alignment of distantly related sequences is delayed, thus overriding the dendrogram. Also thepairwise alignments are performed using local gap penalties and there is automatic selectionand adjustment of the residue substitution matrixand gap penalties, respectively.ClustalW and ClustalX performbest when the sequences to be aligned areglobal cases and have no obvious outlier. Theevolutionary distance between the sequencesmust be relatively low, thus producing a densedendrogram. Also, the penalty scheme used byClustalW/X discriminates against long insertionsand deletions (indels) and will, therefore, exhibit reduced accuracy in such cases. The algorithm is reasonably fast and canhandle sizeable sets of sequences. However, itsspeed decreases when it is given very large setsof data, such as genomic data, and the overallperformance is less accurate when compared tothe other available methods andsuch as POA.

MSA

The simultaneous alignment algorithmMSA employs multidimensionaldynamic programming. Note thatMSA here denotes the name of the Lipman algorithm rather than the abbreviation formultiple sequence alignment used throughoutthis unit. To reduce computations, the MSAmethod employs the Carillo and Lipman [27]approach, which estimates, using pairwisealignments, how much around the *N*-dimensionalsearch matrix diagonal needs to besearched, where *N* is the number of sequencesto be aligned. The Carillo and Lipman methodgeneralizes the earlier pairwise diagonal stripmethod of Ficket to *N* dimensions. Althoughthis approximation of simultaneousalignment optimizes the sum-of-pairs score,which in principle is much more accurate anderrorfree than progressive methods using thensame optimization, it has huge limitations inhow many sequences it can simultaneouslyalign due to its excessive memory and computationalrequirements. Up to 10 sequences of200 to 300 residues in length can be alignedwith the MSA method. The method addressesan additional problem in the comparison ofmultiple sequences, which is the weighting ofthe aligned sequences, as similar sequencesshould not dominate the final alignment.

DCA, OMA

The DCA (Divide-and-Conquer MSA) method [28] is an strict divide-and-conqueralignment algorithm. DCA follows the same strategy as the MSA algorithm by Lipman and performs simultaneous MSA instead of the progressive approach. The DCA approach is an challenge to overcome the computational complexity of the MSA method. The divide-and-conquer strategy first selects the longest sequence in the set to be aligned and cuts it nearits midpoint. The rest of the sequences are also cut at suitable positions, which are calculated through a heuristic method to reduce computational time, and consequently two new subsequence sets arise. This can then be repeated on the subsequence sets until a certain predefined minimum threshold for subsequence length is reached. The smaller the threshold value setting, the faster, but less optimal, the alignment becomes. The now shorter sets of subsequences can then be separately aligned using the MSA algorithm, thus reducing the time and memory requirements. At the end, all the subalignments are concatenated to produce the full final alignment. DCA represents an enhancement in accuracy and speed with respect to MSA, but computational time is still very sensitive to sequence distance and length so that the number of sequences that can be aligned still remains very low. The DCA algorithm has also been implemented as an iterative scheme called OMA. The OMA method denotes an enhancement with respect to speed and accuracy by adding face bounding, gray code enumeration, sequence weighting, realignment of cut positions, and parallelization techniques. The OMA protocol initiates a DCA alignment using a very low sequence length threshold that is only calculated once. Using this calculated threshold, a new larger threshold is produced at each iteration. This means that the alignment becomes slower at each subsequent iteration but also more optimal. The user can set the number of iterations to get a negotiation between alignment speed and quality. Although OMA shows many improvements in memory usage and accuracy, it isstill very computationally demanding for average systems and cannot handle large data sets.

Dialign

Dialign [29] is a local consistencybased alignment algorithm, which, instead of aligning single residues, aligns whole sequence segments. These segments can be imagined as diagonals, as they would appear on a dot plot of a dot matrix analysis. The most recent version, Dialign2 [30]originallyachieves all pairwisealignments of the sequences to be aligned, after which all ungapped segments (diagonals) are identified. Consistent sets of diagonals are then determined and added sequentially to the alignment using an iterative mathematical procedure that determines the optimal order of addition. Only sequence fragments for which matched segments are found are aligned; regions in-between blocks of similar segments are left unaligned. The improvement of Dialign2 compared to Dialign1 is the modification of the original weighting of diagonals, which was formerlybased on Altschul and Erickson. The Dialign2 algorithm is both an accuracy and computational time enhancement over the original method. Morgenstern reported that Dialign2 outperforms many local and global algorithms in identifying related motifs, such as ITERALIGN ClustalW,MultAlin, DCA and Match-Box.Dialign2 has also been shown to perform well in both local and global cases of varying evolutionary distance and, more recently, in multidomain cases against T-Coffee, POA and ClustalW.

MEME

The program MEME [31] is a tool for unsupervised motif searching within DNA and protein sequences, which operates using an expectation maximization (EM) algorithm. It discoveries occurrences of motifs by matching the residue compositionat each position of a putative motif against the general composition of background sequence regions that do not show the motif. Regions viewing the most selective compositions are then selected as motifs. A limitation of the MEME motifs is that they are ungapped, but the program can find multiple occurrences in individual sequences, which on the other hand do not need to be encountered within each input sequence. Another useful feature of the MEME method is that itis geared towards finding DNA palindrome sequences, which are often implicated as DNAbinding sites for proteins

ITERALIGN

The program ITERALIGN [32] is a local iterative algorithm that optimizes the consistency between local pairwise alignments and their embedding in an MSA across all input sequences. It first aligns all ungapped regions of significant local similarity. The highscoring sections are then iterativelychanged by a consensus, based on the distance between them. The prevailing consensus set is used as input to the next round, tillconvergence is reached. Core blocks are minedand optimized using local dynamic programming to further enhance the result. Finally, these blocks are interconnected to produce the final alignment. Each of these aligned blocks canthen be studied autonomously as a potential functional/structural unit.

MACAW

MACAW is the iterative local progressive alignment algorithm, which permits the user to lock or shiftregions in an alignment, while nonlocked subsequences are aligned automatically. The method is semi-automatic and produces blocks of alignments shared by all or a subset of the sequences. It islikely to iteratively define conserved regions such that the fraction of poorly defined segments which must be aligned automatically become fewer at each iteration cycle. The GIBBS technique of Lawrence has been united in the MACAW procedure to detect the local fragments.

Match-Box

This technique Match-Box [33]targets to find ungapped sequence regions with a high degree of similarity across a set of input sequences. This is achieved by matching the frequency distribution of all pairwise aligned sequence fragments, with that derived from shuffled sequences. Using a set of the most similar nine-residue fragments, local alignments are formed for each fragment, if similarity outside a threshold is found withsegments across all other sequences. Boxes ofungapped regions are then defined from these local alignments and accumulated in a final alignment with unaligned amino acids and gaps in between the boxes. Thissystem also produces a reliability index for the aligned positions within the boxes, which trusts on statistics derived from examining a moderately small number of known family alignments.

PileUp

The GCG package alignment program PileUp [34] is a global progressive alignment algorithm. It generates an MSA by means of a simplification of the progressive alignment method of Feng and Doolittle. It creates an UPGMA-based dendrogram and for the alignment of two sets of matched sequences, uses the average alignment similarity score of Corpet. PileUp is limited to 500 sequences, with any single sequence in the final alignment reservedto a maximum length of 7000 characters. If longer sequences are encompassed in the alignment, the number of sequences PileUp can align decreases.

Prrp

Prrp [35] is a global iterative stochastic alignment algorithm. This algorithm is a double-nested approach for MSA optimization. In the inner iteration, the sequences are separated into two groups and subsequently readjustedusing a global group-to-group alignment algorithm. When the inner iteration converges, new pairwise sequence weights are derived from a dendrogram constructed with the UPGMA cluster criterion and used to calculate the alignment scores when sequence blocks are matched. When these weights converge, the outer iteration stops. Gotoh reported betteraccuratenesswhen related to ClustalW. These results were confirmed using JOY, a database of structural alignment and later on BAli-BASE in the assessmentof T-Coffee.

POA

The Partial Order Alignment [36] is an extension of the conventional dynamic programming approach. Instead of performing pairwise alignments following a specific order (from a guide tree), sequencesare aligned in the order in which they are given.The emergent MSA is represented by a “partialorder graph,” in which identical residues within a column are merged and the information of the sequence origin is stored. Thus, despite the shortened representation, all of the information of the MSA is retained. A typical PO-MSA of similar sequences encloses a main “consensus”branch and loops where sequences deviate from each other. The POA dynamic programming matrix reproduces this structure by implementingthe bifurcation points, so that the matrix involvesof multiple two-dimensional layers that part and rejoin according to the PO-MSA graph. The best alignment is found by a conventional trace-back operation. The POA algorithm guarantees that each sequence is aligned to the closest sequence in the growing MSA. POA is a novel local progressive algorithm. The novel feature of this method is that it employs partially ordered graphs to signifyaligned sequences instead of profiles.The progressive strategy for thismethod does not follow a guide dendrogram to decide the order in which the sequences will be aligned, but aligns the input sequences in the order in which they are specified. Each time a new sequence is added to the growing alignment, it is aligned with the most closely related hybrid sequence within the MSA as set by the partial order graph.

PRALINE

PRALINE [37] is a global progressive algorithm. Its unique feature is that it incorporates many policies for alignment optimization. The progressive alignment strategy does not use a precalculated search dendrogram but achieves, at each alignment step, a full profile search with the mostrecently aligned sequence block. It, therefore,re-evaluates, at each alignment step, which sequences or blocks of sequences should be aligned, and hence controls the alignment order during progressive alignment. The pairwise alignments are accomplished using dynamic programming. PRALINE offers a number of policies to optimize the quality of MSA, such as local global alignment and global and local profile preprocessing, and has weighted iteration capabilities. In addition, it can integrate secondary structure to guide the resultant alignment. PRALINE is more of a tool kit than a one-stepMSA method. It allows the user to apply or combine different strategies to a given problem and find the best solution, rather than applying a single approach to everything. The *local global optimization* strategy implements a local alignment of the sequences first, and then, using that information, performs a final global alignment.

SAGA

The program SAGA (Sequence Alignment by Genetic Algorithm) [38] is an iterative stochastic alignment technique that uses a genetic algorithm (GA) [39] to select the alignment froman evolving alignment population and which optimizes, as an Objective Function (OF), the weighted sum of pairs as used in the MSA program. The algorithm primarilycreates a random population of alignments of the sequences, called generation zero (G0). Offspring alignments are then created from the parent alignments in G0 that are estimated for fitness based on alignment superiority. The better the alignment, the more offspring alignments it produces. The operators for offspring alignment creation can be either the mixing of the contents of the parent alignments (crossovers) or the alteration of a single parent (mutation). This process is iterated through successive generations, permittingonly the fittest (best-quality) offspring alignments to advance to the next generation and produce their own offspring alignments. The iteration process halts when no more improvement can be attained. SAGA was found to yield overall better scoring alignments when compared to MSA ofLipman and ClustalW and although optimal alignments could be done for sets of over 30 sequences, the processing time was enormouslyhigh. More recently, a measure of consistency between the considered MSA and a corresponding library of Clustal pairwise alignments was taken.

T-Coffee

T-Coffee (Tree-based Consistency Objective Function For alignmEnt Evaluation) [40] is a global progressive consistency-based algorithm. Initially, all pairwisealignments of the sequences are performedtwice: once with the global alignment method ClustalW where a single global alignment is produced, and once with the local alignment method Lalign where 10 top-scoring nonintersecting local alignments are generated. The results are pooled into aprimary library of pooled weights for each nonredundant residue pair. The pooledweight for each residue pair corresponds to the sum of scores of the global and local alignments containing that residue pair. Each alignment score is the percentage sequence uniqueness of that alignment. A library extension step is then done using a procedure called *matrix extension* tomeasure how residue pairs align with respect to other residues in the library, producing triplet weights. These triplets are then used to assess how well sequences are aligned associated to the other sequences in the data set, rather than looking at pairs of sequences in isolation. The final alignment is builtby performing the library extension step toproduce a guide dendrogram, which then orders how the sequences are aligned. The assessment of Lassmann and Sonnhammer show that T-Coffee is one of the most reliable MSA methods to date in cases of low to moderate evolutionary distance, with higher accuracy compared to ClustalW, Prrp, Dialign2, and POA. In multidomain cases, it shows good performance as a global alignment method and is only outperformed by local alignment strategies. However, T-Coffee has speed and computational demand boundaries when alignments (>30 sequences) of large sequences (>10,000 residues) areachieved and may even fail to complete them on average-powered systems.

MUMmer

Genome-wide sequence alignments needvery fast algorithms that can handle millions of nucleotides. The alignment system MUMmer [41] uses “suffix trees,” which allow for an alignment of two entire genomes in linear time and space. The program finds “maximal unique matches”(MUMs) between two input sequences. A suffix tree is a unique character string where the sequences are equal. A new branch is created where they differ. MUMmer creates a suffix tree based on one (reference) sequence and streams the second (query) sequence against it. MUMmer has been used to assemble contigsfrom shotgun-sequencing to construct the complete genome.

MAFFT

The MAFFT program [42]is centered on the fast Fourier transform (FFT) for speedy detection of homologous segments. Amino acids are denoted by volume and polarity values, yielding signal peaks if homologous segments are aligned. The recorded segments areamalgamated to a final alignment by dynamic programming.

MUSCLE

MUSCLE [43] is anextensively used program. It has attained a higher rank in accuracy and a faster speed compared to ClustalW and T-Coffee. It includes fast distance estimation using kmer counting; progressive alignment using a new profile function called the log-expectation score; and refinement using tree-dependent restricted partitioning. MUSCLE-SMP [44] was the first parallel attempt of MUSCLE on shared memory system. It achieves an overall speedup of 15.2 on a 16 processors SMP system using OperMP. It was combined with the multithreaded algorithm in [45]. It used the bag-of-tasks model. Tests on 16 node cluster showed fascinating improvement for progressive alignments and efficiency scales with the advance in the problem size. MUSCLE-based multiscale simulations [46] have been presented in the two types of infrastructures: local HPC cluster and Amazon AWS cloud solutions. It has been joined with grid space virtual laboratory that permits users to develop and execute virtual experiments on the essential computational and storage resources through its website based interface.

DIALIGN-TX

DIALIGN-TX uses progressive and greedy approaches for segment-based MSA [47]. It integrated anchors optimizations for accurate alignments.

DIALIGN-TX-MPI

It is the parallel version of DIALIGN-TX [48]. It uses an iterative heuristic method for MSA and produces alignments by concatenating ungapped regions with high similarity. It was implemented using both OpenMP and MPI on a 28-cores heterogeneous cluster.

Sample-Align-D

Sample-Align-D [49] is another parallel MSA program. It was based on partitioning the set of sequences into smaller subset using k-mer count based similarity index (k-mer rank). Then each subset is individually aligned in parallel. It has been applied on a cluster of workstation on 16 node using MPI library, and shows a remarkable speedup. It was able to align 2000 randomly selected sequences in less than 10 minutes, compared to over 23 hours on sequential MUSCLE.

MSAProbs

MSAProbs [50] is a new and practical multiple protein sequence alignment algorithm intended by combining a pair-HMM and a partition function to compute posterior probabilities. It also examines two critical bioinformatics techniques, namely weighted probabilistic stability transformation and weighted profile-profile alignment, to achieve high alignment accuracy. In addition, it is optimized for modern multi-core CPUs by employing a multi-threaded design in order to decrease execution time. It statistically proves dramatic accuracy improvements over several top performing aligners.

MSACompro

MSACompro [51] is a new capable and reliable multiple protein sequence alignment program. It includes predicted secondary structure, relative solvent accessibility, and residue-residue contact information into the currently most accurate posterior probability-based MSA methods. It used a multiple-threading implementation on a 32 CPU cores machine. Benchmarks clearly show improvements in accuracy over the leading tools including MSAProbs.

ParaAT

ParaAT [52] is a recent parallel program that is proficient of constructing multiple protein-coding DNA alignments for a large number of homologs. It is well appropriate for large-scale data analysis in the high-throughput era. It allocates each homolog to one of the slave threads; enable user to customize one of multiple sequence aligners (including ClustalW, Mafft, Muscle, T-Coffee); consolidates the results from all slave threads; then parallely back-translates multiple protein sequence alignments into the corresponding DNA alignments. Tests performed on a 24 cores machine provide good scalability and shows high efficiency.

V. CONCLUSION

An MSA can be observedas a representation that offers a unified picture of sequence similarity by averaging out matched residues that perhaps cannot be reliably matched over the entire lengths of the sequences. This is because of evolution, mutations, insertions, and deletions of sequence fragments.So the sequence alignment inconsistencies can well arise under divergent evolution. Given these difficulties, building a reliable MSA for a query set of sequences is a overwhelming task. In this unit it has been made strong that the increased attention to multiple sequence alignment methodology has ensued in recent developments regarding most of its facets. The increased focus has also led to the construction of new benchmark databases and novel evaluation protocols. More developments will be significantlydependent on the integration and representation of biological knowledge in new quality criteria. There are now a multitude of high-quality MSA techniques, each with particular strengths and weaknesses. Increased sensitivity could flourish as a result of new consensus protocols to utilize the combined power of the techniques, or new techniques to determine the kind of alignment problem at hand and then invoke the most appropriate method or combination of methods available. In the meantime, however, it remains important for the end user to run a combination of different MSA methods to optimize the biological information derived from a set of sequences.

REFERENCES

[1] Victor Simossis, Jens Kleinjung, and Jaap Heringa *Current Protocols in Bioinformatics*(2003) 3.7.1-3.7.26Copyright © 2003 by John Wiley & Sons, Inc.

[2] Taylor, W.R. 1988. *A flexible method to align large numbers of biological sequences*. J. Mol. Evol.28:161-169.

[3] Needleman, S.B. and Wunsch, C.D. 1970. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. J.Mol. Biol. 48: 443-453.

[4] Heringa, J. and Taylor, W.R. 1997. *Three-dimensional domain duplication, swapping and stealing*. Curr. Opin. Struct. Biol. 7:416-21.

[5] Heringa, J. 1998. *Detection of internal repeats: How common are they*? Curr. Opin. Struct. Biol.8:338-345.

[6] Smith, T.F. and Waterman, M.S. 1981. *Identification of common molecular sub sequences*. J. Mol. Biol. 147:195-197.

[7] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. 1993. *Detecting subtle sequence signals: A Gibbs sampling strategy for multiple sequence alignment.* Science 262:208-214.

[8] Bailey, T.L. and Elkan, C. 1994. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. pp. 28-36. AAAI Press, Menlo Park, Calif.

[9] Morgenstern, B. 1999. *Dialign 2: Improvement of the segment-to-segment approach to multiple sequence alignment.* Bioinformatics 15:211-218.

[10] Thompson J.D., Higgins, D.G., and Gibson, T.J.1994b. *Improved sensitivity of profile searched through the use of sequence weights and gapexcision.* CABIOS 10: 19-29.

[11] Notredame, C., Higgins, D.G., and Heringa, J. 2000.*T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J. Mol. Biol.302:205-217.

[12] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. *Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.* NucleicAcids Res. 25:3389-3402.

[13] Stoye, J. 1998. *Multiple sequence alignment with the divide-and-conquer method*. Gene 211:GC45-GC56.

[14] Lipman, D.J., Altschul, S.F., and Kececioglu, J.D.1989. *A tool for multiple sequence alignment.*Proc. Natl. Acad. Sci. U.S.A. 86:4412-4415.

[15] Gribskov, M., McLachlan, A.D., and Eisenberg, D.1987. *Profile analysis: Detection of distantly related proteins.* Proc. Natl. Acad. Sci. U.S.A.84:4355-4358.

[16] Hogeweg, P. and Hesper, B. 1984. *The alignment of sets of sequences and the construction of phyletic trees: An integrated method.* J. Mol.Evol. 20:175-186.

[17] Thompson, J.D., Higgins, D.G., and Gibson, T.J.1994a. *CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucl.Acids Res. 22:4673-80.

[18] Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: *A novel method for fast and accurate multiple sequence alignment.* J. Mol. Biol.302:205-217.

[19] Heringa, J. 1999. *Two strategies for sequence comparison: Profile-preprocessed and secondary structure induced multiple sequence alignment.* Comput. Chem. 23: 341-364.

[20] Heringa, J. 2002. *Local weighting schemes for protein multiple sequence alignment.*Comput.Chem. 26:459477.

[21] Feng, D.F. and Doolittle, R.F. 1987. *Progressive sequence alignment as a prerequisite to correct phylogenetic trees.* J. Mol. Evol. 25:351-360.

[22] Hogeweg, P. and Hesper, B. 1984. *The alignment of sets of sequences and the construction of phyletic trees: An integrated method.* J. Mol.Evol. 20:175-186.

[23] Corpet, F. 1988. *Multiple sequence alignment with hierarchical clustering.*Nucl. Acids Res.16:10881-10890.

[24] Taylor, W.R. 1988. *A flexible method to align large numbers of biological sequences*. J. Mol. Evol. 28:161-169.

[25] Barton, G.J. and Sternberg, M.J.E. 1987. *A strategy for the rapid multiple sequence alignment of protein sequences: Confidence levels from tertiary structure comparisons.* J. Mol. Biol. 198:327337.

[26] Higgins, D.G. and Sharp, P.M. 1988. *CLUSTAL: A package for performing multiple sequence alignment on a microcomputer.* Gene 73:237-244.

[27] Carillo, H. and Lipman, D.J. 1988. *The multiple sequence alignment problem in biology.* SIAM J. Appl. Math. 48:1073-1082.

[28] Stoye, J. 1998. *Multiple sequence alignment with the divide-and-conquer method.* Gene 211:GC45-GC56. Stoye, J., Moulton, V.

[29] Morgenstern B., Dress, A., and Werner, T. 1996. *Multiple DNA and protein sequence alignment based on segment-to-segment comparison.* Proc. Natl. Acad. Sci. U.S.A. 93:12098-12103.

[30] Morgenstern, B. 1999. *Dialign 2: Improvement of the segment-to-segment approach to multiple sequence alignment.* Bioinformatics 15:211-218.

[31] Bailey, T.L. and Elkan, C. 1994. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* pp. 28-36. AAAI Press, Menlo Park, Calif.

[32] Brocchieri, L. and Karlin, S. 1998. *A symmetric-iterated multiple sequence alignment of protein sequences.* J. Mol. Biol. 276:249-264.

[33] Deperieux, E. and Feytmans, E. 1992. *Match-Box: A fundamentally new algorithm for the simultaneous alignment of several protein sequences.* Comp. Appl. Biosci. 8:501-509.

[34] GCG. 1993. *Program manual for the GCG Package,* v. 8. Genetics Computer Group, Madison, Wis.

[35] Gotoh, O. 1996. *Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments*. J. Mol. Biol. 264:823-838.

[36] Lee, C., Grasso, C., and Sharlow, M.F. 2002. *Multiple sequence alignment using partial order graphs.* Bioinformatics 18:452-464.

[37] Heringa, J. 1999. *Two strategies for sequence comparison: Profile-preprocessed and secondary structure-induced multiple sequence alignment.* Comput. Chem. 23: 341-364.

[38] Notredame, C. and Higgins, D.G. 1996. *SAGA: Sequence alignment by genetic algorithm*. Nucl. Acids Res. 24: 1515-24.

[39] Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning,* AddisonWesley, New York.

[40] Notredame, C., Higgins, D.G., and Heringa, J. 2000. *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J. Mol. Biol. 302:205-217.

[41] Delcher, A., Phillippy, A., Carlton, J., and Salzberg, S.L. 2002. *Fast algorithms for large-scale genome alignment and comparison.* Nucleic Acids Res. 30:2478-2483.

[42] Katoh, K., Misawa, K., Kuma, K., and Miyata, T. 2002. *MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res. 30:3059-3066.

[43] R. C. Edgar, *MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids* Res., vol. 32, No. 5, 2004, pp. 1792-1797.

[44] Xi. Deng, , E. Li, J. Shan and W. Chen, *Parallel implementation and performance characterization of muscle,* IPDPS ’06, 2006, pp. 1-7.

[45] Evandro A. Marucci et al., *Using Threads to Overcome Synchronization Delays in Parallel Multiple Progressive Alignment Algorithms,* American Journal of Bioinformatics, vol. 1, No. 1, 2012, pp. 50-63.

[46] K. Rycerz et al., *Comparison of Cloud and Local HPC Approach for MUSCLE-based Multiscale Simulations,* e-ScienceW ’11, 2011, pp. 81-88.

[47] A. R. Subramanian, M. Kaufmann, and B. Morgenstern, *Dialign-TX: Greedy and Progressive Approaches for Segment-Based Multiple Sequence Alignments,* Algorithm Mol. Biol., vol. 3, No. 6, 2008, doi:10.1186/17487188-3-6.

[48] E. de Araujo Macedo et al., *Hybrid MPI/OpenMP Strategy for Biological Multiple Sequence Alignment with DIALIGN-TX in Heterogeneous Multicore Clusters*, IPDPSW ’11, 2011, pp. 418-425.

[49] F. Saeed, and A. Khokhar, *Sample-Align-D: A High Performance Multiple Sequence Alignment System using Phylogenetic Sampling and Domain Decomposition*, IPDPS ’11, 2008, pp. 1-9.

[50] Yongchao Liu, Bertil Schmidt, Douglas L. Maskell, *MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities,* Bioinformatics, vol. 26, No. 16, 2010, pp. 1958 -196.

[51] X. Deng and J. Cheng, *MSACompro: Protein Multiple Sequence Alignment Using Predicted Secondary Structure, Solvent Accessibility, and Residue-Residue Contacts,* BMC Bioinformatics, vol. 12, 2011, pp. 472488.

[52] Z. Zhang, J. Xiao, J. Wu, H. Zhang, G. Liu, X. Wang, and L. Dai, *ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments*, Biochemical and Biophysical Research Communications, vol. 419, No. 4, 2012, pp. 779-781.