Journal of Computer Science 5 (9): 651-660, 2009
ISSN 1549-3636

Genome Sequence Analysis: A Survey

Hassan Mathkour and Muneer Ahmad
Department of Computer Science, College of Computer and Information Sciences
King Saud University, P.O. Box 51178, Riyadh 11543
Kingdom of Saudi Arabia

Abstract: Problem statement: Sequence analysis problems are NP hard and need optimal solutions. Interesting problems include duplicate sequence detection, sequence matching by relevance, sequence analysis using approximate comparison in general or using tools i.e., Matlab and multi-lingual sequence analysis. The usefulness of these operations is highlighted and future expectations are described. **Approach:** This study described the concepts, tools, methodologies, algorithms being used for sequence analysis. The sequences contained precious information that needed to be mined for useful purposes. There was high concentration required to model the optimal solution. The similarity and alignments concepts can not be addressed directly with one technique or algorithm, a better performance was achieved by the comprehension of different concepts. **Results:** We had compared different approaches using exemplary data and found that ClustalW2 is fairly good tool in terms of analysis. We assigned different weight values for relevant features and obtained score 95 in comparison phenomenon and 45 in alignment. **Conclusion:** Different techniques and approaches had been evaluated and compared.

Key words: Genome, multi-lingual, approximate matching, nucleotide base pair, corpora, duplicate sequences

INTRODUCTION

Sequences are logical units that contain vital information, for instance consider biological sequences that compose of nucleotide base pairs in the form of A (Adenine), T (Thymine), G(Guanine) and C (cytosine). The structure and position of these pairs in sequence determine the personality, habits and inheritance characteristics of species.

The mining of useful information from the vast repositories of sequence data brings interesting results related to genes and their functional properties, the main attention and focus of biologists is to differentiate species on behalf of these functional characteristics, many different solutions have been proposed that claim to bring optimal results. It is worth knowing that direct matching in sequence repository data is not efficient and may bring inaccurate and slow results, so going beyond the exact match is necessary for optimality.

Modern computational technology and good devices has made the job of scientists relatively easy in bringing accurate results, this reflection is quite positive in micro-array DNA technology and image data-sets comparison techniques where huge bulky genetic data is approximately compared promptly.

The data is spread over chips and relevancy is determined. The other tools like MATLAB, TRADOES and EBMT are now broadly used for sequence manipulation. FASTA and BLAST are also very popular in biological researchers for sequence comparisons, different people have developed many tools for analysis of not only the genetic sequences but corpora sequences, the lexical analysis explores the hidden resources in these structures, global alignment tools have replaced local one and multiple alignment techniques have given way to know more about diversity in functional properties of species in sequences.

People are interested in mining some kind of association rules in genetic and lexical data, these rules will better help to understand the patterns in data and further exploration may lead to more knowledgeable and interesting results that could not be available by query application phenomenon. The query application only generates views that are provided through datasets within a confined domain and redefined rules in the form of queries, later solutions present the query enhancement techniques but that are not as optimal as direct rule generation from datasets.

Corresponding Author: Muneer Ahmad, Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Riyadh Saudi Arabia

Scientists now use latest systems in biotechnology for storage of genetic data, employing data-ware housing techniques and analyzing the DNA sequences, it is not limited to computations but can solve many different complex biological problems.

The more comprehensive use of these computer aided techniques falls in field of molecular medicine, which is itself a broad filed that involve physical, biological and chemical methods for depiction of molecular structure. Another important aspect of genome analysis is building evolutionary models and phylo-genetic tree structure

Fig. 1 describes the sequence analysis hierarchy. In this hierarchy, at the top, general sequence analysis depicts that sequence may be of different nature and kinds, for instance, genetic sequences, protein sequences, multi language sequences, corpora and other mathematical set of occurrences of events or characters or symbols. In genome sequence analysis, biologists are paying a very keen attention to the alignment and micro array analysis today as alignment leads towards interesting facts regarding diversity in species, genetic relationship between species and degree of relevancy that how much one creature is different and similar from others. The micro-array technology brings very collective and near results for sequence analysis and is thought to be a future technology.

MATERIALS AND MATHODS

Sequence comparison: Sequence comparison is a kind of method in which two or more than two sequences are chosen for searching for certain domain specific patterns that need an alignment procedure at first glance, for instance, bioinformatics people quote two kinds of alignment, local and global.

Local is a kind of point to point alignment while global alignment on other hand is spread to a more concentrated area of search which may involve search at different regions, e.g., (Fig. 2).

Sequence analysis tools: Following are a few tools developed for sequence analysis.

EMBOSS: This tool has been developed to compared two sequences, it has two sections/parts, one is called Needle which is used when comparison is required at whole length of both sequences and other is called Water which provides region wise similarity in strands.

CLUSTALW 2: It provides good meaningful sequence match for both DNA and protein sequences and separately shows the degree of similarity and differences in strands in a kind of visual environment and also provides an evolutionary relationship between sequences.

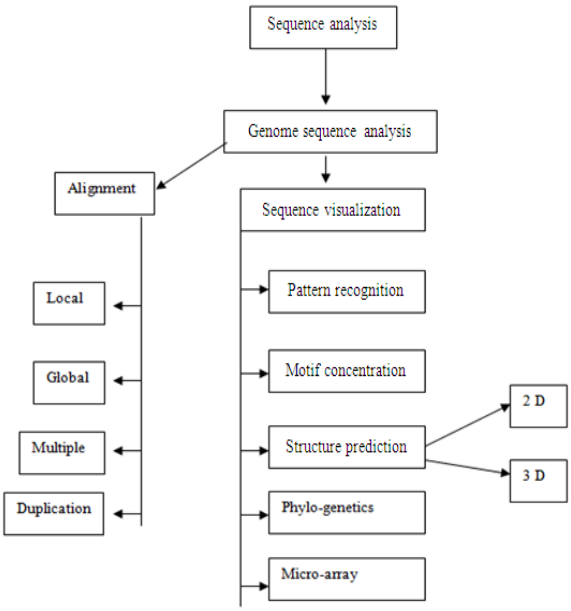

Fig. 1: Sequence analysis hierarchy

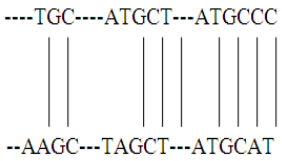

Fig. 2: Global alignment

KALIGN: It is supposed to be a fast and accurate multi sequence alignment tool. It requires a supported format for input data strands and can also input data by user command lines, it can provide interactive sequence results for both protein and DNA strands.

MAFFT: MAFFT is a tool designed for alignment of sequences using Fast Fourier Transforms, it is claimed to be high level multi alignment tool with prompt and quick results. The beauty of this tool is the GAP extension feature provided and also requires a specific format for input data strands.

MUSCLE: MUSCLE is a multi sequence alignment tool and compares the sequence by LOG EXPECTATION; it is supposed to provide better performance than CLUSTALW2 or T-COFEE, it also requires strands to be in specific format and can generate out put data tree that fan help better understand the alignment.

T-COFFEE: It is also a multi sequence alignment program that has the capability to combine the

alignment being derived from some other alignment programs, so it provides a kind of refinement from other tools, it can produce the alignment in a sequence of two by two resulting in global and local alignment.

The phase-wise alignment can be then combined in an integrated final refined multi alignment structure.

Exemplary comparison of tools: Suppose we have a genetic data file that contains sequences of human and mouse.

The genetic data for mouse is sequenced[22] as:

>FOSB_MOUSE Protein fosB

MFQAFPGDYD SGSRCSSSPS AESQYLSSVD
SFGSPPTAAA SQECAGLGEM PGSFVPTVTA
ITTSQDLQWL VQPTLISSMA QSQGQPLASQ
PPAVDPYDMP GTSYSTPGLS AYSTGGASGS
GGPSTSTTTS GPVSARPARA RPRRPREETL
TPEEEEKRRV RRERNKLAAA KCRNRRRELT
LPGSTSAKED GFGWLLPPPP PPPLPFQSSR
DAPPNLTASL FTHSEVQVLG DPFPVVSPSY
TSSFVLTCPE VSAFAGAQRT SGSEQPSDPL
NSPSLLAL

And human sequence[22] is:

>FOSB_HUMAN Protein fosB

MFQAFPGDYD SGSRCSSSPS AESQYLSSVD
SFGSPPTAAA SQECAGLGEM PGSFVPTVTA
ITTSQDLQWL VQPTLISSMA QSQGQPLASQ
PPVVDPYDMP GTSYSTPGMS GYSSGGASGS
GGPSTSGTTS GPGPARPARA RPRRPREETL
TPEEEEKRRV RRERNKLAAA KCRNRRRELT
LEFVLVAHKP GCKIPYEEGP GPGPLAEVRD
LPGSAPAKED GFSWLLPPPP PPPLPFQTSQ
DAPPNLTASL FTHSEVQVLG DPFPVVNPSY
TSSFVLTCPE VSAFAGAQRT SGSDQPSDPL
NSPSLLAL

Running the specimen data on EBI CLUSTALW2, the alignment score is 2031 and both sequences contain same no. of characters.

From this discussion, it is obvious that pair-wise alignment of both human and mouse genomes have been shown with representation of symbol (*) where match is found and symbol (.) where characters are mismatched, the overall score is 95 for both sequence pairs.

Executing the same data set for EBI Align, we get the gap penalty 10 and extension penalty as 0.5.

The sequence lengths are same, identity representation is 95.9% and similarity is 97.6%, gaps and score are 0.0% and 1693 respectively, the similarity representation is done by vertical lines and difference is shown by (.).

MAFFT takes input of both the strands and keeps some default gap penalty, the gap extension is set to 0.123 and gap open are set to 1.53.

Kalign builds the sequence gpo to 11.0 and gpe to 0.85, the alignment is not shown in the form of symbols and clear identifiers are not made so that one has to pay more concentration while viewing the alignment visuals but it is considered to be much better then MAFFT.

MUSCLE (with same datasets execution) generates no gap penalty and gap extension; rather it shows the alignment similarity and differences in the form of visual colors. T-COFFEE generates an alignment score of 61 without mentioning the gap penalty and gap extension, it also displays the results aligned with the introduction of symbol (*) for similarity and (.) for difference.

Table 1 depicts the comparative analysis of various tools run on same dataset, the difference in results shows that each tool has tried to solve the NP hard problem of sequence alignment with diverse context, some tools have generated visual alignment and others have given numerical scores.

Table 2 shows different criteria's in terms of various features of tools. There is a scoring scheme for measurement of cumulative performance of each tool. Local, global and multiple alignments have been given weight 0.15 each out of 1.

Table 1: Comparative analysis of tools

Tool	Alignment score	Gap penalty	Extension penalty	Identity
CLUSTANW2	2031	10	0.5	95
Align	1693	0.0	0.0	95.9
MAFFT	Not shown	Default	0.123	Not shown
KALIGN	Not shown	11	0.85	Not shown
MUSCLE	Not shown	Nil	Nil	Not shown
T-COFFEE	Not shown	Nil	Nil	61

Table 2: Comparative analysis in terms of features

Criteria	CLUSTALW2	Align	MAFFT	KALIGN	MUSCLE	T-COFFEE
Local Alignment (15 %)	2*0.15 3*0.15	3*0.15	1*0.15	1*0.15	1*0.15	
Global Alignment (15%)	2*0.15	3*0.15	3*0.15	1*0.15	1*0.15	1*0.15
Multiple alignment (15%)	3*0.15	1*0.15	3*0.15	3*0.15	3*0.15	3*0.15
Visual Depiction (20 %)	3*0.2	3*0.2	3*0.2	3*0.2	3*0.2	3*0.2
General Score Results (15%)	3*0.15	3*0.15	1*0.15	1*0.15	1*0.15	3*0.15
Phylo-genetic Tree (10%)	3*0.1	1*0.1	1*0.1	1*0.1	1*0.1	1*0.1
GAP consideration (10%)	3*0.1	3*0.1	3*0.1	3*0.1	3*0.1	2*0.1
Total (Sum*100/6)	45	42	42	31	31	35

Visual representation of alignment is also a strong feature of a certain alignment tool that has been given weight 0.2. Similarly general score and Phylo-genetic tree depiction are weighted as 0.15 and 0.1. Another important feature for an alignment measuring tool is gap consideration which is given weight 0.1. Against each tool, a percentage sum over sum of (1 = absent, 2 = average, 3 = present) is calculated. Table 2 shows that CLUSTALW2 is well performed tool that contains average features for local and global alignment, full features for multiple alignment, full features for visual representation of strands, full features for score and gap consideration and a very powerful feature that is Phylo-genetic tree representation of specimen aligned data, this feature is lacking in all other tools which makes CLUSTALW2 much significant as compared to others.

A review of previous work: Bansal[1] presents a considerable useful idea in the form of a frame work that treats multiple sequences as abstract data type and integrates the information gathered from this frame work. The information gathered is helpful for generation of phylo-genetic tree. Authors have developed a generic high level language library for complex analysis of multiple sequences and derived groups of amino acids in homologous protein which share some common properties along with identification of constrained columns which also conserve some common properties despite mutations resulting into different types of amino acids in the column. PROLOG TOOL is being used to be applied on proposed frame work. A high level abstraction is used at alignment of sequences with the introduction of prolog tool, which some times is not quite useful for generating standard optimal results and overall comparison is not quite visible[1].

Kappen[2] described an annotated technique for comparisons between a mouse chromosome 9 and a human chromosome 15, the data draft sequences had been obtained from genetic databases and a complex map containing 14 genes has been presented as a genome map, the framework described in the study for data interpretation and demonstration can be quite helpful for generation of more complex maps provided time constrained is kept in mind, the ideas may lead towards implementation of automated genome annotation techniques[2]. A very useful feature of this approach was to use information for human and mouse species comparatively and to describe a frame work for the discovery of three previously unknown genes. The limitation of this framework requires more labor in the form of critical evaluation before accepting any kind of predictions and focus must be made on smaller region in the map to bring more sophisticated results[2].

Nahar[3] presented a web based tool that provides comparative genome sequence analysis, this tool is interactive and user of the program can interact with different parts to view/monitor better results. The claim is that idea is novel and one may not analyze DNA sequence directly but with the help of self adjusting maps that could provide possible evolutionary concepts in depiction of certain results. The authors described a strong advantage of this tool to be highly interactive in visual identification of horizontally transferred genes and this kind of functionality is not available in other techniques/tools. The weakness of idea is that some time the user is not intending with the maps and eager to get final approximate results with ease without interacting with application interfaces, secondly the tool is web based so actual application complexity in the form of time frame may not be possible.

Chang[4] proposes a package of integrated comparative analysis for comparison of different genomes, the framework develops efficient gene identification and functional annotations and plots numerous measures for all positions in a long DNA sequence and can perform whole genome comparison[4]. The framework proposes a cross-species pathway comparison on customized starting and ending points of pathways. The idea is comprehensively good as it can depict both section-wise and whole complete analysis, the example illustrated in the study does not cover or highlight the complete idea and more sophisticated understanding may reveal the hidden aspects[4]. More analysis tools for comparative analysis may be required. Cornell[5] has proposed a data-ware house (Genome Information Management System GIMS) that incorporated both genomic sequence and functional data[5]. This ware house has been explained by giving an example of yeast genome data. It can answer many useful queries and serves as a basis for future exploration by creating a large data-ware house with genomic and functional features. The claim is that this framework will provide better effective analysis of genome with functional properties and focuses the development of data management and analysis techniques for use with multiple genome data-sets. If comprehensive storage is available then genomic data-ware housing is good appealing idea that can replace conventional approaches for genomic analysis[5]. A little weakness is that more efforts and work is required for the construction of genomic warehouse.

Ahmed[6] has proposed an algorithm that is experimentally evaluated in a distributed grid environment that provides very scalable and low computational cost[6]. As multiple sequence alignment and comparison problem falls in a domain of length so parallel approach focusing on the parts of sequences and then integrated can lead to better approximate results, so main focus remains on utilization of grid computing for large biological data. The algorithm was studies in three different distributed environments including a single cluster environment, a single cluster grid environment and a multi cluster grid environment[6]. A distributed environment is essentially required for application of this approach with many more addition of resources which may be costly as compared to traditional approaches.

Agrawal[7] proposes a heuristic approach for multiple sequence alignment. The author claims that dynamic programming algorithm involves computational complexity that brings slow and inefficient results, the author compares proposed algorithm with CLUSTALW which takes O(N2n2) time and claims that modified technique works for O (N log2(Nn2)), the proposed approach also makes the alignment process more dynamic as the order of sequences added to the multiple sequence alignment also depends on the already computed multiple sequence alignment[7]. The claim is not supported with examples and results; more study is required to depict some solid fruitful results.

Cai[8] has described a comprehensive evolutionary computational approach for multiple sequence alignment by representing a set of 17 clusters of orthologous groups of proteins and compared the results with the standard results from CLUSTALW and found the proposed results better than the standard approach[8]. One major weakness of the idea is that current implementation uses the fixed parameter tractable algorithm for gap 0-1 alignments, it is not feasible for finding alignments when the number of sequences is much larger than 15. The comparison is quite good for small scale and not efficient for large scaled sequences.

Liu Weiguo[9] proposed a streaming approach for multiple sequence alignment, this approach is based on PC graphics hardware, using modern graphics processing units for high performance computing with low cost make it possible to depict more sophisticated results, the authors have reformulated dynamic programming algorithm bases alignment as streaming algorithm in terms of computer graphics hardware boundaries. The proposed technique is quite comprehensively efficient with only weakness of system graphics hardware primitives. Suitable graphics hardware is mandatory for application and execution of approach.

Zhao[10] presents an improved Ant Colony algorithm that is more sophisticated form of previous technique, the authors claim that their modified approach can operate genomic sequences of any length while traditional Ant Colony approach uses fixed length sequences, the modified approach can avoid local optimum problem, so proposed technique brings robust and efficient results. The weakness of this approach is that searching small chunk in larger sequences may bring bad or erroneous results which may reveal the fact that using this approach for multiple sequences alignment would not be so useful as compared to traditional approaches[10].

Arslan[11] described an improved algorithm for multiple sequence alignment problem, this approach considers two layers each of which corresponds to part of the dynamic programming matrix for the alignment of the given sequences and computes each layer differently using dynamic programming technique, in this way the proposed approach is much more efficient than traditional approach that uses weighted automata and performance is claimed to be much better than other approaches.

Davidson [12] depicts an approach that is basically an integration of dynamic programming and heuristic approach with minimal amount of additional overhead, the idea is that dynamic matrix is traversed along anti-diagonals, bounding the computation to exclude partitions of the matrix that can't contain optimal paths, so the heuristic approach will prune the unnecessary paths from this matrix and present an optimal solution to the problem[12]. The second benefit of this approach is that it presents an efficient use of memory by using divide and conquers technique at the cost of some system computations, the weakness of this approach is that implementing for an arbitrary dimensional matrix will be much more difficult than a two dimensional case. Secondly more dissimilar sequences can bring bad results.

Rashid[13] shows a fast dynamic programming based sequence alignment algorithm uses the reduced amino acids alphabet to transform the protein sequences into a sequence of integers and uses n-gram to reduce the length of the sequence and then traditional approach is used to get the similarity measure between two sequences[13]. The results of this proposed approach seem to be quite satisfactory as compared to traditional approaches. Another benefit of this approach is that it

requires less space then traditional approaches as it shortens the length of sequences each time but computational overhead is also involved.

Agrawal[14] claims a better performance by presenting a modification to the iterative approach by incorporating in it the use of multiple parameter sets. Preliminary experiments indicate that using multiple parameter sets gives significantly better performance than using a single parameter set and than using a simple match/mismatch scoring scheme. The authors generate a family of matrices at various distances and multiple matrices for different conservation rates have been used for bringing an optimal alignment. The only weakness of this approach is that using too many parameters may degrade performance.

RESULTS AND DISCUSSION

Following techniques serve as foundation for building blocks regarding comparative genome sequence analysis,

The methods are discussed below and their comparative analysis is presented in Table 3 and 4:

- Dynamic programming method as an extension
- Progressive methods
- Iterative methods

Dynamic programming methods as an extension: The dynamic programming method[29,30] used for Global Alignment of a pair of sequences can be extended for Multiple Sequence Alignment. But the limitation of this method is that it can not efficiently align more sequences, when the no. of sequences grows, the performance of the method degrades considerably.

Progressive methods: Progressive Methods[28] use the Dynamic Programming Method to built the MSA (Multiple Sequence Alignment) starting with most related sequences and then progressively adding less related sequences to initial alignment. e.g.:

- CLUSTALW
- PILEUP

The drawbacks of Progressive Methods are dependent of initial pair-wise Sequence Alignment. The very first sequences must be very closely related sequences, if sequences are closely aligned then there will be few errors but if sequences are not closely aligned there will be more errors.

Iterative methods of MSA: Iterative Methods[29] attempt to correct for the problem raised by Progressive Methods by repeatedly realigning subgroups of sequences and then by aligning these subgroups into Global Alignment[29,30]. The programs MultiAlin and DIALIGN align multiple sequences using these methods[30].

Table 3: Performance comparison of methods

Method	Approach	Applicability	Suitability	Non suitability	Performance
Dynamic programming	Attempts to match all pairs in sequences and builds a scoring scheme	Sequence alignment	Local and global alignment	Multi alignment	Good for local and global but involves much computational overhead
DP as an extension	Extension of DPA for global alignment	Sequence alignment	All kinds of alignments	Lengthy strands of DNA and protein	It is an extension to DPA but restricted to small and medium strands, degrade with increasing size of chains
Progressive methods	DPA based and align most relevant sequence and then grow incrementally	Sequence alignment	Multiple sequence alignment	Non suitable for sequences having much initial dissimilarity	Appreciable for strands with initial similarity and degrades with diverse chains
Iterative methods	Based on progressive methods and resolve problems raised in realigning	Sequence alignment	Multiple sequence alignment	Lengthy and initial dissimilar chains	Appreciable for small and initial similar strands involves computational overhead in realigning

Table 4: Comparison between approaches

Category	Proposed Approach	Strengths	Weaknesses	Reference No.
Sequence analysis	Multi sequence analysis for generation of Phylo-genetic tree	Novel idea for generation of phylo -genetic tree, development of high level language library, complex analysis performed	Overall comparison performance is not visible and high level abstraction does not bring feasible results with PROLOG	[1,24,30,34]
Comparative analysis	Genetic comparison between human and mouse genomes	Presented genome maps that can be quite useful for comprehension of complex maps with limited constrained, discovered three unknown genes	Limitation requires more labor in critical evaluation, focus should be made on smaller regions in map	[2,25,36]
Comparative analysis	Web based tool for comparative genome sequence analysis	Interactive tool for ease of user, introduced self adjusting maps for visualization	User may not be intended with maps and wish to get some approximate final results, web based tool may not reflect time complexity accurately	[3,4,23,27]
Sequence analysis	Genome sequence analysis tool for visualization	Depicts efficient gene identification and functional annotations, performs whole genome comparison	Need more illustration of idea rather than an example to reveal the hidden truths	[3,20,34]
Data storage and retrieval	Proposed an approach for building genome data-ware house	Incorporates both genomic and functional data, can provide better effective analysis of genetic data	Efforts and labor involves in construction of huge genomic sequence data with lot of memory requirement and computation	[5,33,35]
Sequence alignment	Multi sequence alignment with restricted domain	Proposed approach provides scalable and efficient results in distributed grid environment with lower computation cost	Grid computing with distributed environment may require more resources that cost than traditional approaches	[8,11,15,19]
Sequence alignment	Proposed a heuristic approach for multi sequence alignment	Approach is more good than CLUSTALW2 and there is considerable efficiency by lower running time, the alignment process is more dynamic	Require comprehensive illustration with examples and results, more work is required to get some solid results	[7,11,14,18]
Sequence alignment	Multi sequence alignment with restricted domain	Computational approach for MSA with set of clusters brings more good results than CLUSTALW2	With the increase of more multi sequences, the performance will degrade, is feasible for small domain only	[11,12,19]
Sequence alignment	Streaming approach for multi sequence alignment	Researchers have reformed the dynamic programming algorithm and used streaming approach for better results, it is claimed to be highly efficient with introduction of graphic hardware	System graphic hardware primitives, suitable graphic hardware is mandatory for implementing idea	[10,13,16]
Sequence alignment	An improved ant colony algorithm for multi sequence alignment	Can operate on any genomic sequence of any length, avoids local optimum problem and brings efficient and robust results	Finding small set of sequence data in large data set may slow down the system and may bring faulty results	[10,15]
Sequence alignment	An improved algorithm for multi sequence alignment	Takes help from dynamic programming matrix, much more efficient than traditional approach that uses some kind of weighted automata	Initially a better choice is mandatory for sequence of motifs otherwise results will be difficult to get accurate	[7,12,19]
Sequence alignment	An integrated approach for multi sequence alignment	Proposed an idea to traverse dynamic matrix anti- diagonally with avoidance of non optimal paths, manage memory efficiently by divide and conquer technique	Implementation of an arbitrary dynamic matrix is more difficult than an ordinary two dimension matrix	[8,9,13]
Sequence alignment	Sequence alignment algorithm based on fast dynamic programming algorithm	Requires less space and brings more good results than traditional approach, it is also a kind of integration of two approaches so hybrid is always considered suitable	In this hybrid approach, more computational overhead is involved	[7,9,11,12,13, 16,17]
Sequence alignment	Multi sequence alignment using multi parameter sets	Multi parameter sets give more performance than single parameter one, the scoring schemes and different distance matrices can bring better results	Too many parameters may degrade the performance	[8,14,18]
Sequence alignment	Multi sequence alignment using portioned optimization algorithm	Improves solution time and quality, layered approach brings good results, avoids local optimal traps	Involves more overhead in the form of computational stuff	[15]

Table 4: continue

Sequence alignment	Solution of sequence alignment problem	Require less time and space to solve the sequence analysis problem, is also suitable for other local and global sequence optimization problems, can handle both small and large sequences	Internal calculation are complex and computational stuff is involved	[13,17,31,36]
Sequence alignment	Multi sequence alignment using fuzzy logic	Enhances the performance of genetic algorithm, the probability of three operations of genetic algorithm are quite fast and accurate and align sequences more efficiently	Local search may degrade the system performance, scoring matrix and space scores concept are more traditional	[18,29]
Sequence alignment	Sequence matching using fuzzy logic	The assembler designed can work with low quality data, the performance measures of assembler were found accurate than other assemblers	Relies on enhanced fuzzy logic technique and fuzzy approximate methodology	[17,18,32]
Sequence alignment	Multi sequence alignment using recursive technique	Can operate on set of sequences with local, global and multi alignment, recursive in nature, certain degree of performance can be evaluated at all levels	Dividing the system into smaller blocks can bring computational overhead, the local alignment phenomenon should not be addresses with multiple one	[19,31]
Sequence alignment	Sequence comparison using Matlab histogram comparison	Novel idea for sequence comparison, Matlab brings accurate results	May require more space for pixel calculation and image comparisons	[20,32]
Sequence alignment	Duplicate sequence detection	Genetic databases may contain redundant sequence information, the algorithm can overcome redundant sequence structure	Require more time in sequence pattern matching due to huge size of genetic data	[21]

CONCLUSION

Bioinformatics is a very rapidly emerging field of research; the genome sequence analysis is a very interesting and challenging task that needs great attention and focus. The analysis brings very promising relevance between species. We are now able to find certain genetic similarity and differences in apparently different and diverse creatures, the micro-array technology, phylo-genetic tree creation and many other alignment and analysis tools have helped biologist greatly.

Future expectations: The genome sequence analysis will help biologist to devise genetic therapy and solutions for genetic disorders. It will also open ways to explore genetic diversity in species; a very challenging goal of this study will be to uncover the wealth of biological information hidden in genetic data. A good generalization of these concepts will better help in areas of molecular medicines that would provide more generic sophisticated medicines for curing diseases. It is definitely a genomic revolution and next decade will reveal the real work and achievement for biologists.

REFERENCES

1. Bansal, A.K., 1995. Establishing a framework for comparative analysis of genome sequences. Proceeding of the 1st International Symposium on Intelligence in Neural and Biological Systems, May 29-31, IEEE Computer Society, Washington DC., USA., pp: 84-91. http://portal.acm.org/citation.cfm?id=854059
2. Kappen, C. and J.M. Salbaum, 2003. Comparative genome annotation for mapping, prediction and discovery of genes. Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Jan. 6-9, IEEE Computer Society, Washington DC., USA., pp: 1-275. http://portal.acm.org/citation.cfm?id=821789
3. Nahar, N., L. Hamel, M.S. Popstova and J.P. Gogarten, 2007. GPX: A tool for the exploration and visualization of genome evolution. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, Oct. 14-17, IEEE Xplore Press, Boston, MA., pp: 1338-1342. DOI: 10.1109/BIBE.2007.4375743
4. Chang, Y.F., C.Y. Chen, H.W. Chen, I.H. Lin, W.X. Luo, C.H. Yang, Y.H. Lin and C.H. Chang, 2005. Bioinformatics analysis for genome design and synthetic biology. Proceeding of the Emerging Information Technology Conference, Aug. 15-16, IEEE Xplore Press, USA., pp: 2. DOI: 10.1109/EITC.2005.1544358
5. Cornell, M., N.W. Paton, W. Shengli, C.A. Goble and C.J. Miller *et al*., 2001. GIMS-a data warehouse for storage and analysis of genome sequence and functional data. Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering Conference, Nov. 4-6, IEEE Xplore Press, Bethesda, MD., USA., pp: 15-22. DOI: 10.1109/BIBE.2001.974407

6. Ahmed, N.Y. Pan and A. Vandenberg, 2005. Parallel algorithm for multiple genome alignment on the Grid environment. Proceedings of the 19th IEEE International Symposium on Parallel and Distributed Processing, Apr. 4-8, IEEE Xplore Press, USA., pp: 7. DOI: 10.1109/IPDPS.2005.324
7. Agrawal, A. and S.K. Khaitan, 2008. A new heuristic for multiple sequence alignment, Proceedings of the IEEE International Conference Electro/Information Technology, May 18-20, IEEE Xplore Press, Ames, IA., pp: 215-217. DOI: 10.1109/EIT.2008.4554299
8. Cai, L.D. Juedes and E. Liakhovitch, 2000. Evolutionary computation techniques for multiple sequence alignment. Proceedings of the 2000 Congress on Evolutionary Computation, July 16-19, IEEE Xplore Press, La Jolla, CA, USA., pp: 829-835. DOI: 10.1109/CEC.2000.870716
9. Liu Weiguo, B. Schmidt, G. Voss and W. Muller-Wittig, 2007. Streaming algorithms for biological sequence alignment on GPUs. IEEE. Trans. Paral. Distribut. Syst., 18: 1270-1281. http://portal.acm.org/citation.cfm?id=1313072
10. Yidan, Z., M. Ping, L. Jie, L. Chun and J. Guoli, 2008. An improved ant colony algorithm for DNA sequence alignment. Proceedings of the International Symposium on Information Science And Engineering, Dec. 20-22, IEEE Computer Society, USA., pp: 683-688. http://www2.computer.org/portal/web/csdl/doi/10.1109/ISISE.2008.82
11. Arslan, A.N. and D. He, 2006. An improved algorithm for the regular expression constrained multiple sequence alignment problems. Proceedings of the 6th IEEE Symposium on Bioinformatics and Bioengineering, Oct. 16-18, IEEE Computer Society, Washington DC., USA., pp: 121-126. http://portal.acm.org/citation.cfm?id=1169392
12. Davidson, A., 2001. A fast pruning algorithm for optimal sequence alignment. Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering Conference, May 4-6, IEEE Computer Society, Washington DC., USA., pp: 49-56. http://portal.acm.org/citation.cfm?id=791300
13. Rashid, N.A.A., R. Abdullah, A.Z.H. Talib and Z. Ali, 2006. Fast dynamic programming based sequence alignment algorithm. Proceedings of the 2nd International Conference on Distributed Frameworks for Multimedia Applications, May 2006, IEEE Xplore Press, Pulau Pinang, pp: 1-7. DOI: 10.1109/DFMA.2006.296909
14. Agrawal, Ankit, Huang and Xiaoqiu, 2008. Pairwise DNA alignment with sequence specific transition-transversion ratio using multiple parameter sets. Proceedings of the International Conference on Information Technology, Oct. 17-20, IEEE Xplore Press, Bhubaneswar, pp: 89-93. DOI: 10.1109/ICIT.2008.62
15. The European Bioinformatics Institute for research in system biology and bioinformatics, HTTP://www.ebi.ac.uk/Tools/clustalw2/alignment.txt
16. Sung, W.K. and W.H., Lee, 2003. Fast and accurate probe selection algorithm for large genome. Proceedings of the IEEE Computer Society Conference on Bioinformatics, Aug. 4-11, IEEE Computer Society, Stanford, California, pp: 65. http://www2.computer.org/portal/web/csdl/doi/10.1109/CSB.2003.1227305
17. Le, S.Y., J.H. Chen and J.V. Maizel, 2003. Statistical inference for well-ordered structure in nucleotide sequence. Proceedings of IEEE Computer Society Conference on Bioinformatics, Aug. 4-11, IEEE Computer Society, Stanford, California, pp: 190. http://www2.computer.org/portal/web/csdl/doi/10.1109/CSB.2003.1227318
18. Zavolan, D.M., N. Socci, N. Rajewsky and T. Gaasterland, 2003. SMASHing regulatory sites in DNA by Human-mouse sequence comparisons. Proceedings of IEEE Computer Society Conference on Bioinformatics, Aug. 4-11, IEEE Computer Society, Stanford, California, pp: 277-286. http://www2.computer.org/portal/web/csdl/doi/10.1109/CSB.2003.1227328
19. Greely, H.T., 2001. Genotype Discrimination: The complex case for some legislative protection. Univ. PA. Law Rev., 149: 1483-1505. http://www.ncbi.nlm.nih.gov/pubmed/15732207
20. John Wagner and Phyllis Gardner, 1997. Towards cystic fibrosis gene therapy. Ann. Rev. Med., 48: 203-216. http://direct.bl.uk/bld/PlaceOrder.do?UIN=025547780&ETOC=RN&from=searchengine
21. Agrawal, R., C. Faloutsos and A. N. Swami, 1993. Efficient similarity search in sequence databases. Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms, Oct. 1315, IEEE Computer Society, Washington DC., USA., pp: 69-84. http://portal.acm.org/citation.cfm?id=645415.652239

22. Amitay, E., R. Nelken, W. Niblack and R. Sivan, 2003. Multi-resolution disambiguation of term occurrences. Proceedings of the 12th Conference on Information and Knowledge Management, Nov. 3-8, IEEE Computer Society, New Orleans, LA., USA., pp: 255-262. http://portal.acm.org/citation.cfm?id=956863.956913
23. Artiles, J., A. Penas and F. Verdejo, 2004. Word sense disambiguation based on term to term similarity in a context space. Proceedings of the 3rd International Workshop on Evaluation of Systems for Semantic Analysis of Text, Association for Computational Linguistics, 2004, pages 58-63
24. Baeza-Yates, R.A. and G.H. Gonnet, 1999. A fast algorithm on average for all against-all sequence matching. Proceedings of the International Workshop and Symposium on String Processing and Information Retrieval, Sept. 22-24, IEEE Xplore Press, Cancun, Mexico, pp: 16-23. DOI: 10.1109/SPIRE.1999.796573
25. Baeza-Yates R.A. and G. Navarro, 1996. A faster algorithm for approximate string matching. Proceedings of 7th Combinatorial Pattern Matching Symposium, July 10-12, Springer-Verlag, London, UK., pp: 1-23. http://portal.acm.org/citation.cfm?id=738447
26. Baldwin, T. and H. Tanaka, 2000. The effects of word order and segmentation on translation retrieval performance. Proceedings of the 18th International Conference on Computational Linguistics, July 31-Aug. 4, IEEE Computer Society, New Orleans, LA., USA., pp: 35-41. http://portal.acm.org/citation.cfm?id=990826
27. Banerjee, S. and T. Pedersen, 2003. Extended gloss overlaps as a measure of semantic relatedness. Proceedings of 18th International Joint Conference on Artificial Intelligence, (IJCAI'03), Lawrence Erlbaum Associates Ltd., USA., pp: 805-810. http://direct.bl.uk/bld/PlaceOrder.do?UIN=146297376&ETOC=RN&from=searchengine