₹ 200 ISSN - 2249-555X
Volume : 1 Issue : 7 April 2012

Indian Journal Of Applied Research

Journal for All Subjects

www.ijar.in
Listed in International ISSN Directory, Paris.

ISSN - 2249-555X

Indian Journal of Applied Research

Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College, New Congres Bhavan, Paldi,
Ahmedabad-380006, Gujarat, INDIA
Contact.: +91-9824097643 E-mail : editor@ijar.in

INDEX

Volume : 1 | Issue : 7 | April 2012 ISSN - 2249-555X

Research Paper Engineering

Multiple Sequence Alignment of Different Species

* Prerna ** Pankaj Bhambri ***Dr. O.P. Gupta

* Research Scholar, Department of C.S.E, G.N.D.E.C., Ludhiana

** Assistant Prof., Department of Info. Tech, G.N.D.E.C., Ludhiana

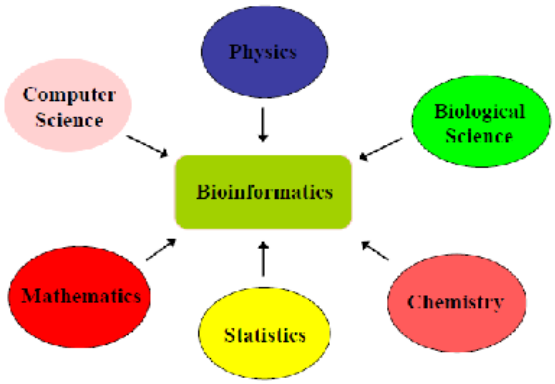
*** Dy. Director, School of Info. Tech, P.A.U., Ludhiana

Keywords :

1. Introduction to Bio- informatics

The term bioinformatics was developed by Pauline Hogeweg and Ben Hesper for the study of informatics processes in bio molecules. Bioinformatics is the application of computer science and information technology to the field of biology and medicine. There are the many programming languages that used in this field are Java, C, C++,MatLab.

Fig 1. Bioinformatics



2. PROBLEM FORMULATION

The formulation of the problem reaches its end when the finished work is being obtained. The major problem being faced by today's researchers and biologists is huge amount of raw data and how to store and effectively use this data. The analysis of multiple sequence alignment is to align number of sequences to calculate their matching scores. Based on the scores sequences are defined that are closely related and distantly related.

There are various methods and web based programs used to perform multiple sequence alignment. Pair-wise alignments may suffice to create links between structure and function. MSA are very powerful because two sequences that may not align well to each other can be aligned via their relationship to a third sequence, thereby integrating information in a way not possible using only pair-wise alignments. The most common web based method is ClustalW. ClustalW produces the best match for the selected sequences, and arranges them so that the identities, similarities and differences can be seen.

The steps are used to perform ClustalW:

1. Perform pair-wise alignment for all sequences.
2. Use the alignment scores, that gives a phylogenetic tree.
3. The sequences are progressively aligned using the phylogenetic relationship indicated by the tree.

One of the most important features of this program is the flexibility of using substitution matrices. Clustal does not rely on a single substitution matrix. Instead, it applies different scoring matrices when aligning sequences, depending on degrees of similarity. The choice of a matrix depends on the evolutionary distances measured from the guide tree.

3. METHODOLOGY

Sequence alignment is the task of identifying structural or evolutionary relationship between the collected sequences of amino acids. Modern programs that are used for constructing MSA usually consist of two components: obejective function for accessing the quality of alignment and optimization procedure for identifying the highest scoring alignment.

Practically dynamic programming is not feasible for 3 or more sequences, so heuristic approaches have been developed. One of the methods of heuristic approach is progressive alignment. Progressive alignment works by aligning most alike sequences then adding less related sequences to alignment procedure. CLUSTALW is the main program of progressive alignment. The alignment problem of long and comparatively different sequences is very complex. ClustalW results the best match for the selected sequences, and arranges them according to their identities, similarities and differences.

Some MATLAB functions are also used to align multiple sequences. There are list of functions to read the information about molecules, retrieve the information from the online database sites like NCBI, EBI, align the sequences, read the fasta format of the sequence etc. Scoring matrices are calculated using PAM and BLOSUM methods. These methods have predefined values to calculate scores according to match, mismatches and gaps.

Example of ClustalW :

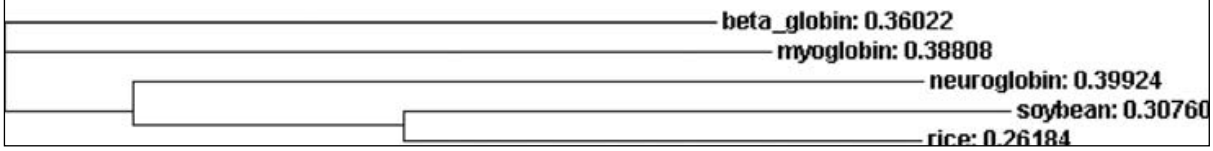
Stage 1: generate a series of pairwise alignments

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43

1

Stage 2: create a guide tree, calculated from a distance matrix

```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```



Stage 3: Alignment of sequences

beta globin ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin ----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin ----------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59

beta globin DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS----ELHCDKLHVDPE 102
myoglobin HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA----QSHATKHKIPVK 103
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS--LGRKHRAVGVKLS 104
soybean --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA---LGSVHAQKAVTDP 101
rice --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117

beta globin NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166

Figure 3.1: ClustalW Analysis

4. MULTIPLE SEQUENCE ALIGNMENT

The purpose of multiple sequence alignment algorithms is to detect evolutionary, and thus structural and functional, relations among sequences. The successful sequence comparison would allow us to infer the biological properties of new sequences from data accumulated on related genes. In trying to characterize a newly discovered sequence, the first step is always to check whether there are similar sequences in known databases and their annotation.

A reasonable colour scheme is:

Table 1.1 Color Scheme

Colour	Residue type	Amino acids
Yellow	Small Nonpolar	Gly, Ala, Ser, Thr
Green	Hydrophobic	Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, Trp
Magenta	Polar	Asn, Gln, His
Red	Negatively Charged	Asp, Glu
Blue	Positively charged	Lys, Arg

4.1 Scoring Matrices
The different scoring matrices used for homology prediction are PAM and BLOSUM.

4.1.1 Point Accepted Mutations (PAM)
PAM is the point mutation per 100 amino acids. PAM1 means a 1 point mutation/100 amino acids. A PAM matrix is usually a matrix of 20 by 20 which gives the different rates of mutations between pairs of amino acids. One of the first amino acid **substitution matrices**, the PAM matrix was developed by Margaret Dayhoff. This matrix is calculated by observing the distances in closely related proteins. The PAM1 matrix estimates what rate of substitution would be expected if 1% of the amino acids had changed. Using this logic, Dayhoff derived matrices as high as PAM250. Usually the PAM 30 and the PAM70 are used.

Table 4.1 PAM70 Substitution Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	5	-4	-2	-1	-4	-2	-1	0	-4	-2	-4	-4	-3	-6	0	1	1	-9	-5	-1	-1	-1	-2
R	-4	8	-3	-6	-5	0	-5	-6	0	-3	-6	2	-2	-7	-2	-1	-4	0	-7	-5	-4	-2	-3
N	-2	-3	6	3	-7	-1	0	-1	1	-3	-5	0	-5	-6	-3	1	0	-6	-3	-5	5	-1	-2
D	-1	-6	3	6	-9	0	3	-1	-1	-5	-8	-2	-7	-10	-4	-1	-2	-10	-7	-5	5	2	-3
C	-4	-5	-7	-9	9	-9	-9	-6	-5	-4	-10	-9	-9	-8	-5	-1	-5	-11	-2	-4	-8	-9	-6
Q	-2	0	-1	0	-9	7	2	-4	2	-5	-3	-1	-2	-9	-1	-3	-3	-8	-8	-4	-1	5	-2
E	-1	-5	0	3	-9	2	6	-2	-2	-4	-6	-2	-4	-9	-3	-2	-3	-11	-6	-4	2	5	-3
G	0	-6	-1	-1	-6	-4	-2	6	-6	-6	-7	-5	-6	-7	-3	0	-3	-10	-9	-3	-1	-3	-3
H	-4	0	1	-1	-5	2	-2	-6	8	-6	-4	-3	-6	-4	-2	-3	-4	-5	-1	-4	0	1	-3
I	-2	-3	-3	-5	-4	-5	-4	-6	-6	7	1	-4	1	0	-5	-4	-1	-9	-4	3	-4	-4	-3
L	-4	-6	-5	-8	-10	-3	-6	-7	-4	1	6	-5	2	-1	-5	-6	-4	-4	-4	0	-6	-4	-4
K	-4	2	0	-2	-9	-1	-2	-5	-3	-4	-5	6	0	-9	-4	-2	-1	-7	-7	-6	-1	-2	-3
M	-3	-2	-5	-7	-9	-2	-4	-6	-6	1	2	0	10	-2	-5	-3	-2	-8	-7	0	-6	-3	-3
F	-6	-7	-6	-10	-8	-9	-9	-7	-4	0	-1	-9	-2	8	-7	-4	-6	-2	4	-5	-7	-9	-5
P	0	-2	-3	-4	-5	-1	-3	-3	-2	-5	-5	-4	-5	-7	7	0	-2	-9	-9	-3	-4	-2	-3
S	1	-1	1	-1	-1	-3	-2	0	-3	-4	-6	-2	-3	-4	0	5	2	-3	-5	-3	0	-2	-1
T	1	-4	0	-2	-5	-3	-3	-3	-4	-1	-4	-1	-2	-6	-2	2	6	-8	-4	-1	-1	-3	-2
W	-9	0	-6	-10	-11	-8	-11	-10	-5	-9	-4	-7	-8	-2	-9	-3	-8	13	-3	-10	-7	-10	-7
Y	-5	-7	-3	-7	-2	-8	-6	-9	-1	-4	-4	-7	-7	4	-9	-5	-4	-3	9	-5	-4	-7	-5
V	-1	-5	-5	-5	-4	-4	-4	-3	-4	3	0	-6	0	-5	-3	-3	-1	-10	-5	6	-5	-4	-2
B	-1	-4	5	5	-8	-1	2	-1	0	-4	-6	-1	-6	-7	-4	0	-1	-7	-4	-5	5	1	-2
Z	-1	-2	-1	2	-9	5	5	-3	1	-4	-4	-2	-3	-9	-2	-2	-3	-10	-7	-4	1	5	-3
X	-2	-3	-2	-3	-6	-2	-3	-3	-3	-3	-4	-3	-3	-5	-3	-1	-2	-7	-5	-2	-2	-3	-3

4.1.2 BLOSUM
Dayhoff's methodology of comparing closely related species turned out not to work very well for aligning evolutionarily divergent sequences. Changes to sequence over long evolutionary time scales are not well approximated by compounding small changes that occur over short time scales. The BLOSUM series of matrices rectifies this problem. The method of probabilities used in the matrix calculation is computed by looking at "blocks" of conserved sequences found in multiple protein alignments. These conserved sequences are assumed to be of functional importance within related proteins. The BLOSUM75 matrix is calculated from observed substitutions between proteins that share 75% sequence identity.

Table 4.2 BLOSUM75 Substitution Matrix

A	R	N	D	CQ	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R-	2	6	-1	-2	-4	1	0	-3	0	-3	-3	2	-2	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-5
N-	2-	1	6	1	-3	0	-1	-1	0	-4	-4	0	-3	-4	-3	0	0	-4	-3	-3	3	0	-1	-5
D-	2-	2	1	6	-4	-1	1	-2	-1	-4	-4	-1	-4	-4	-2	-1	-1	-5	-4	-4	4	1	-2	-5
C-	1-	4-	3-	4	9	-3	-5	-3	-4	-1	-2	-4	-2	-2	-4	-1	-1	-3	-3	-1	-4	-4	-2	-5
Q-	1	1	0	-1	-3	6	2	-2	1	-3	-3	1	0	-4	-2	0	-1	-2	-2	-2	0	3	-1	-5
E-	10		-1	1	-5	2	5	-3	0	-4	-4	1	-2	-4	-1	0	-1	-4	-3	-3	1	4	-1	-5
G	0	-3	-1	-2	-3	-2	-3	6	-2	-5	-4	-2	-3	-4	-3	-1	-2	-3	-4	-4	-1	-2	-2	-5
H-	2	0	0	-1	-4	1	0	-2	8	-4	-3	-1	-2	-2	-2	-1	-2	-2	2	-4	-1	0	-1	-5
I-	2-	3-	4-	4-	1-	3-	4-	5-	4	4	1	-3	1	0	-3	-3	-1	-3	-2	3	-4	-4	-2	-5
L-	2-	3-	4-	4-	2-	3-	4-	4-	3	1	4	-3	2	0	-3	-3	-2	-2	-1	1	-4	-3	-1	-5
K-	1	2	0	-1	-4	1	1	-2	-1	-3	-3	5	-2	-4	-1	0	-1	-4	-2	-3	-1	1	-1	-5
M-	1-	2-	3-	4-	2	0	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-2	1	-3	-2	-1	-5
F-	3-	3-	4-	4-	2-	4-	4-	4-	2	0	0	-4	0	6	-4	-3	-2	1	3	-1	-4	-4	-2	-5
P-	1-	2-	3-	2-	4-	2-	1-	3-	2-	3-	3-	1-	3-	4	8	-1	-1	-5	-4	-3	-2	-2	-2	-5
S	1	-1	0	-1	-1	0	0	-1	-1	-3	-3	0	-2	-3	-1	5	1	-3	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-2	-1	-1	-2	-1	1	5	-3	-2	0	-1	-1	-1	-5
W-	3-	3-	4-	5-	3-	2-	4-	3-	2-	3-	2-	4-	2	1	-5	-3	-3	11	2	-3	-5	-3	-3	-5
Y-	2-	2-	3-	4-	3-	2-	3-	4	2	-2	-1	-2	-2	3	-4	-2	-2	2	7	-2	-3	-3	-2	-5
V	0	-3	-3	-4	-1	-2	-3	-4	-4	31		-3	1	-1	-3	-2	0	-3	-2	4	-4	-3	-1	-5
B-	2-	1	3	4	-4	0	1	-1	-1	-4	-4	-1	-3	-4	-2	0	-1	-5	-3	-4	4	0	-2	-5
Z-	1	0	0	1	-4	3	4	-2	0	-4	-3	1	-2	-4	-2	0	-1	-3	-3	-3	0	4	-1	-5
X-	1-	1-	1-	2-	2-	1-	1-	2-	1-	2-	1-	1-	1-	2-	2-	1-	1-	3-	2-	1-	2-	1-	1-	5
*-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5-	5	1

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The model is constructed for predicting different proteins sequence alignments (or DNA) along with the alignment scores. The protein alignment is useful in predicting the homologous sequences and thus useful for assigning the class for unknown protein. There are different sequence formats available from which plain text format is utilized. Multiple alignment uses the fasta format to align the sequences along column. Scoring matrix assists in computing the alignment score. It consists of mutation scores for different amino acid alignments. The model is user friendly which provides different options to the user. The overall computation is also shown in the matrix along with the final sequence alignment.

5.2 Future Scope of Work

Following improvements regarding the developed model of bioinformatics can be made:

- Alignment of the sequences of different species can be incorporated.
- Protein structure prediction can be incorporated to further enhance the model.
- Database can be maintained for the known protein sequences and the search option can be provided to the user.
- Different scoring matrices can be included.

6. SCREEN SHOTS

Figure 6.1: Input of the sequence accession numbers

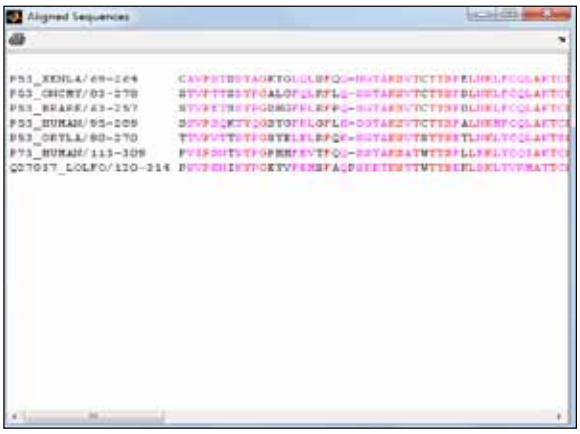


Figure 6.2: Result of Aligned Sequences

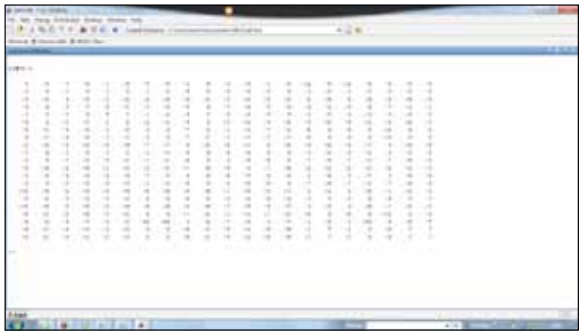


Figure 6.3: Result of PAM Matrices

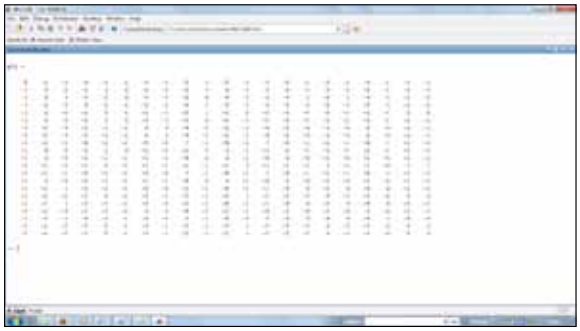


Figure 6.4: Result of BLOSUM Matrices

REFERENCES

[1] Luscombe, N., M., Greenbaum, D. and Gerstein, M. (2001),"Yearbook of Medical Bioinformatics", Department of Molecular Biophysics and Biochemistry Yale University New Haven, USA, pp. 83-100. | [2] Murphy, J. and Walz J. (2011), "Multiple Sequence Alignment : CSL", pp. 1-12. | [3] Edgar, R. C. and Batzoglou S. (2006), "Multiple Sequence Alignment", Current option in structural Biology 2006, pp 16: 368-373. | [4] Rastogi, S., C., Mendiratta. N. and Rastogi, P. (2007) "Genomics, Proteomics and Drug Discovery", Prentice-Hall of India Private Limited, New Delhi, pp.6-25. | [5] Azim, G., A., Othman, M., B. and Eleneen, Z., A. (2011) "Modified Progressive Strategy for Multiple Proteins Sequence Alignment" international journal of computers issue 2, volume 5 pp. 270- 280. | [6] Agarwal, A. and Khaitan, S., K. (2008) "A New Heuristic for Multiple Sequence Alignment" 978-1-4244-2030-8/08/ © IEEE, pp. 215-217. | [7] Johnson, R. (2002) "MATLAB Programming Style Guidelines" pp. 1-14. | [8] Xiong, J. "Essential Bioinformatics" Cambridge University Press (2006) pp. 63-70. | [9] Pevsner, J. "Bioinformatics and Functional Genomics" Wiley- BlackWell Publication (2009) pp. 179-190.

Sara Publishing Academy
Indian Journal Of Applied Research
Journal for All Subjects

Editor,
Indian Journal Of Applied Research
8-A, Banans, Opp. SLU Girls College,
New Congres Bhavan, Paldi, Ahmedabad-380006.
Contact.: +91-9824097643 E-mail : editor@ijar.in

Printed at Unique Offset, Novatsing Rupam Estate, Opp. Abhay Estate, Tavdipura, Shahibaug, Ahmedabad