Optimal Distance Matrix for Multiple Alignment of Amino acid sequences

Ayako OHSHIRO[†*], **Takeo OKAZAKI**[††],

[†]Graduate School of Medicine, University of the Ryukyus
[††]Faculty of Engineering, University of the Ryukyus
1 Senbaru, Nishihara-cho, Okinawa 903-0213, Japan
*Correspondence

Summary
□ Sequence alignment programs comprise numerous algorithms including scoring matrices, e.g., the distance matrix, and methods of elucidating relationships among different target amino acid sequences. Since amino acid sequences have various biological properties, an optimal combination of distance matrices and methods should be selected. This study aimed to identify an optimal distance matrix for each biological characteristic from a plurality of distance matrix prepared in advance for the amino acid sequences acquired from database site with comparative experiments.

Key words:
Distance Matrix, Amino acid sequence, Alignment

1. Introduction

In current bioinformatics studies, analysis of genetic information has gained increasing importance. Sequence alignment has allowed for the elucidation of evolutionary relationships and the estimation of biological functions via alignment of two or more sequences based on the similarity ratio between residues, referred to as a distance matrix. Adequate data are available to establish evolutionary distances among gene sequences. Recently, many alignment algorithms have been proposed, and a comparative analysis of each alignment algorithm has been performed. Hirosawa et al. [1] investigated the performance of different iterative algorithms. Wallace et al. [2] systematically assessed several different iterative algorithms by comparing the results regarding sets of alignment test cases, using HOMSTRAD database of structure-based alignments [3]. Wakatsu [4] performed a comparative analysis to identify the characteristics of different types of datasets and alignment strategies. Here, numerous scoring matrices indicate the evolutionary distance between amino acid residues. For that reason, the choice of distance matrix influences the results of alignment. Hence, it is important to clarify how the alignment results vary depending on the distance matrix. Herein, we performed a comparative experiment of alignment result by changing distance matrix for each biological property.

2. Multiple Alignment with a Distance Matrix

Alignment of two sequences is called pairwise alignment [5][6]; more than three sequences, multiple alignment. Multiple alignment comprises various algorithms, the most commonly used one being a progressive alignment algorithm following a heuristic approach to align numerous sequences. The following algorithm and Fig. 1 show the procedure for progressive alignment.

Step1 Conduct pairwise alignment for all combinations of sequences

Step2 Based on pairwise alignment scores, cluster the sequences into groups in descending order of scores, and construct the guide tree.

Step3 Conduct progressive alignment on the basis of each guide tree

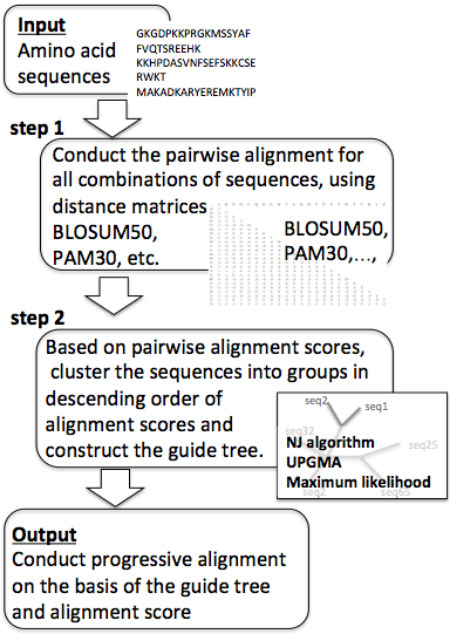

Input
Amino acid sequences
GKGDPKKPRGKMSSYAF
FVQTSREEHK
KKHPDASVNFSEFSKKCSE
RWKT
MAKADKARYEREMKTYIP
step 1
Conduct the pairwise alignment for all combinations of sequences, using distance matrices
BLOSUM50, PAM30, etc.
BLOSUM50, PAM30,...,
step 2
Based on pairwise alignment scores, cluster the sequences into groups in descending order of alignment scores and construct the guide tree.
seq2
seq1
seq32
NJ algorithm
UPGMA
Maximum likelihood
seq25
seq2
seq65
Output
Conduct progressive alignment on the basis of the guide tree and alignment score

Fig. 1 A schematic representation of progressive alignment

Manuscript received October 5, 2018
Manuscript revised October 20, 2018

In **step 1**, the distance matrix is required for pairwise alignment. The distance matrix was first generated by Steven Henikoff and Jorja Henikoff [5], and currently used distance matrices include BLOSUM50, PAM30, etc. In **step 2**, there are some variations in algorithms for the construction of guide trees, e.g., the NJ algorithm, maximum likelihood method, etc. According to these combinations of distance matrices and algorithms for construction of guide trees, various progressive alignments can be undertaken. The present study focused on variations in distance matrices.

Mutation Data (MD) score is based on the concept of the Point Accepted Mutation (PAM). An evolutionary distance of 1 PAM indicates the probability of a residue undergoing a mutation during a distance wherein one point mutation is accepted per 100 residues. The amino acid residues are ranked and grouped here in accordance with their physicochemical properties. For example, sequences clustered at greater than or equal to 80% identity are used to generate the BLOSUM80 matrix (BLOcks SUbstitution Matrix pronounced blossom); those in the 50% or greater cluster contributing to the BLOSUM50 matrix, etc.

Table 1: Mutation Data for BLOSUM50

A	5																			
R	-2	7																		
N	-1	-1	7																	
D	-2	-2	2	8																
C	-1	-4	-2	-4	13															
Q	-1	1	0	0	-3	7														
E	-1	0	0	2	-3	2	6													
G	0	-3	0	-1	-3	-2	-3	8												
H	-2	0	1	-1	-3	1	0	-2	10											
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5										
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5									
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6								
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7							
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8						
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10					
S	1	-1	1	0	-1	0	-1	-1	-1	-3	-3	0	-2	-3	-1	5				
T	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-1	-1	-2	-1	2	5			
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15		
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	3	-1	5
	A	R	N	F	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

For example, to align two sequences "YHER" and "CHKR," using distance matrix BLOSUM50 shown in Table 1, the alignment score between each pair of residues must be determined (Fig. 2).

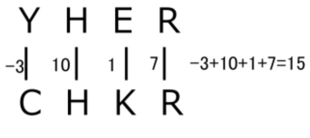


Fig. 2 core calculation

The alignment score of "Y" and "C" is -3, that of "H" and "H" is 10, etc.; thus, the sequence alignment score is determined to be 15.

3. Determination of the optimal distance matrix

To determine the optimal distance matrix, three conditions may be considered. First, the optimal distance matrix has to yield an alignment that is close to the true alignment. Hence, benchmark databases are required for true alignment. Second, since biological sequences are of different types, it is difficult to determine the best suited distance matrix; hence, the optimal distance matrix for each biological category should be determined. Third, since many sequences exist in each biological feature, a distribution of scores is obtained. The average is hence considered the representative index, and a high average is considered favorable. Similarly, variance is also a representative index for measuring the stability of data; lower the variance, better the alignment. Considering these conditions, we proposed the following procedure to determine the optimal distance matrix for multiple alignment.

Step1 Prepare the benchmark database whose true alignment is known

Step2 Classify amino acid sequences on the basis of biological characteristics

Step3 Prepare distance matrices for comparative experiments

Step4 Determine the alignment score with each distance matrix for whole sequences

Step5 Select a distance matrix with a high average and low variance as the optimal distance matrix

From the aforementioned procedure, an optimal distance matrix for each biological feature can be expected to be determined.

4. Experimental Environment

Experimental data, evaluation value, and distance matrices are to be considered for comparative experiments. We compared the following 9 distance matrices: 6 BLOSUM matrices including BLOSUM45, BLOSUM50, BLOSUM60, BLOSUM62, BLOSUM80, and BLOSUM90, and 3 PAM matrices including PAM30, PAM70, and PAM250. BAliBASE [8] is an amino acid database of manually refined multiple sequence alignments specially designed to evaluate and compare multiple sequence alignment programs comprising 218 reference alignments in total, divided into six different reference sets, each with different characteristics (Table 2).

Table 2: Data characteristics of BaliBASE

References	Sets	Contents
11	38	equidistant sequence (very divergent)
12	44	equidistant sequence (medium to divergent)
2	41	evolutionarily distant sequences
3	30	subgroups with a residue identity of <25% between groups
4	49	sequences inserted long gap with terminal

We focused on changes in the alignment results based onthe distance matrix. In the same multiple alignment algorithm, only the scoring matrix is changed. The difference between true alignment and the results of alignment would form the basis of the comparison. For example, using one of the BAliBASE data comprising 8 sequences has a length of 96. They are medium to divergent. We compared the scoring matrices of BLOSUM50 and BLOSUM62 (Table 3). SP score depends on the distance matrix; hence, the SP score is naturally expected to change. BAliBASE SP is the distance to true alignment. When the BAliBASE score is 1, alignment is completely correct.

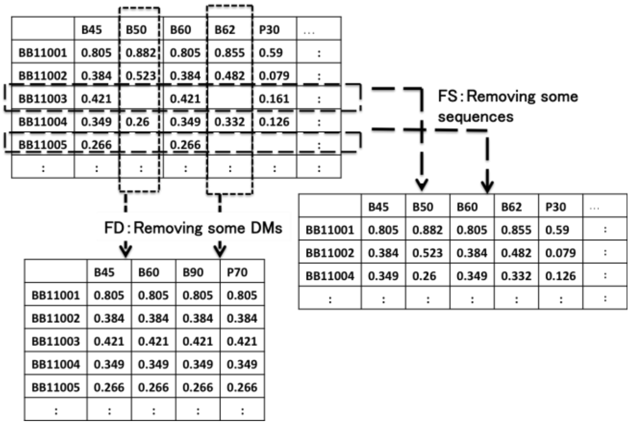
Table 3: Comparison of BLOSUM50 and BLOSUM62

	BAliBASE SP score
BLOSUM50	0.523
BLOSUM62	0.482

In this case, the BAliBASE score of BLOSUM50 is higher than that of BLOSUM62; hence, BLOSUM50 is useful for these sequences.

In the BALiBASE database, each sequence has a biological description. We selected keywords from the descriptions, e.g., for sequences, the description "Aldehyde dehydrogenase" is observed. Hence, those sequences are associated with the feature of dehydrogenation. Thus, 15 features were derived, e.g., Protein, Enzyme Degradative Enzyme, Synthesis enzyme, The post hydrogen enzyme, Phosphorylation enzyme, Transcriptase enzyme, Catalyze enzyme, Intravital material, Molecule, Compound, Bound region, Component of medicine, Virus, and Amino acid.

However, since some distance matrices do not have a score for amino acids sequences, some distance matrices cannot calculate BAliBASE scores for comparative analysis. To solve this program, we made two types of complete datasets, eliminating some sequences with a null score, termed FS, and eliminating some distance matrices with a null score, termed FD.



	B45	B50	B60	B62	P30	…
BB11001	0.805	0.882	0.805	0.855	0.59	:
BB11002	0.384	0.523	0.384	0.482	0.079	:
BB11003	0.421		0.421		0.161	:
BB11004	0.349	0.26	0.349	0.332	0.126	:
BB11005	0.266		0.266			:
:	:	:	:	:	:	:

FS : Removing some sequences

	B45	B50	B60	B62	P30	…
BB11001	0.805	0.882	0.805	0.855	0.59	:
BB11002	0.384	0.523	0.384	0.482	0.079	:
BB11004	0.349	0.26	0.349	0.332	0.126	:
:	:	:	:	:	:	:

FD : Removing some DMs

	B45	B60	B90	P70
BB11001	0.805	0.805	0.805	0.805
BB11002	0.384	0.384	0.384	0.384
BB11003	0.421	0.421	0.421	0.421
BB11004	0.349	0.349	0.349	0.349
BB11005	0.266	0.266	0.266	0.266
:	:	:	:	:

Fig. 2 Data modifying

For the FS and FD datasets, the top three optimal distance matrices with a high average and low dispersions for each biological property are shown in Tables 4 and 5. "DM," "Ave," "Dis," and "RV" denote the optimal distance matrix, average, dispersion, and reference type sequences. The representation "N" implies that the variable value is not calculated because data are not adequate, and "NA" indicates results that are not determined owing to the possible elimination of all sequences during the data modification step.

Table 4: Results of the FS dataset

	1				2				3			
Attribute	DM	Ave	Dis	RV	DM	Ave	Dis	RV	DM	Ave	Dis	RV
Protein	B50	0.712	0.230	12	B80	0.690	0.180	12	B62	0.696	0.230	12
Enzyme	B62	0.732	0.160	12	B50	0.729	0.147	12	B80	0.714	0.157	12
Degradative Enzyme	B50	0.699	0.148	12	B62	0.693	0.150	12	B80	0.674	0.160	12
Synthesis enzyme	N	N	N	N	N	N	N	N	N	N	N	N
The post hydrogen enzyme	B50	0.740	N	11	N	N	N	N	N	N	N	N
Phosphorylation enzyme	B62	0.734	0.240	12	B80	0.734	0.211	12	B50	0.721	0.200	12
Transcriptase enzyme	NA	NA	NA	N	NA	NA	NA	N	N	N	N	N
Catalyze enzyme	NA	NA	NA	N	NA	NA	NA	N	N	N	N	N
Intravital material	P250	0.802	0.050	N	B62	0.795	0.050	12	B50	0.790	0.06	12
Molecule	N	N	N	N	N	N	N	N	N	N	N	N
Compound	N	N	N	N	N	N	N	N	N	N	N	N
Bound region	N	N	N	N	N	N	N	N	N	N	N	N
Component of medicine	N	N	N	N	N	N	N	N	N	N	N	N
Virus	B50	0.810	0.060	12	P250	0.783	0.060	12	N	N	N	N
Amino acid	N	N	N	N	N	N	N		N	N	N	N

Table 5: Results of the FD dataset

	1				2				3			
Attribute	DM	Ave	Dis	RV	DM	Ave	Dis	RV	DM	Ave	Dis	RV
Protein	**B80**	0.704	0.160	12	**P70**	0.698	0.160	12	B90	0.660	0.160	12
Enzyme	**B80**	0.810	0.090	2	B90	0.806	0.090	2	B45, B60	0.806	0.100	2
Degradative Enzyme	**B80**	0.735	0.140	12	B90	0.724	0.140	12	N	N	N	N
Synthesis enzyme	**B80**	0.812	0.020	12	B90	0.810	0.020	12	B60	0.810	0.030	12
The post hydrogen enzyme	P70	0.670	0.200	11	N	N	N	N	N	N	N	N
Phosphorylation enzyme	**B80**	0.790	0.170	12	B90	0.770	0.170	12	N	N	N	N
Transcriptase enzyme	**B90**	0.780	0.120	2	B45, B60	0.780	0.130	2	N	N	N	N
Catalyze enzyme	**B45, B60**	0.802	0.078	2	B90	0.802	0.080	2	N	N	N	N
Intravital material	B90	0.750	0.129	2	B45, B60	0.740	0.135	20	N	N	N	N
Molecule	P70	0.580	0.170	11	N	N	N	N	N	N	N	N
Compound	B90	0.643	0.105	2	B45, B60	0.643	0.116	20	N	N	N	N
Bound region	N	N	N	N	N	N	N	N	N	N	N	N
Component of medicine	**B45**	0.844	0.060	12	P70	0.830	0.020	12	B80	0.814	0.02	12
Virus	B45, B60	0.775	0.079	12	B90	0.775	0.106	12	N	N	N	N
Amino acid	N	N	N	N	N	N	N	N	N	N	N	N

Upon comparing the results of the two datasets, many results of data with close evolutionary distance were obtained. PAM30 and PAM70 are not ranked in Table 4; PAM30 and PAM250, Table 5. BLOSUM50 and BLOSUM80 have a high average and a low dispersion for protein, enzyme, degradative enzyme, synthesis enzyme, and phosphorylation enzyme from Table 4; however, BLOSUM50 is not observed in Table 5. This implies that since BLOSUM50 was excluded during data modification for the generation of the FD dataset, BLOSUM80 is more robust than BLOSUM50. Furthermore, for Molecule, Compound, and Bound region, scores have been determined for the FS but not FD datasets, implying that usable distance matrices are limited; hence, their sequence structures are predicted to be complex.

In addition, to confirm the unsuitable distance matrix for each biological characteristic, results displaying low averages and high dispersions for the FS and FD datasets are shown in Tables 6 and 7.

PAM30 and PAM70 have low scores for Protein and Enzyme; however, PAM70 has a high score and low dispersion for Protein with RV12 (Table 5).

Table 6: Results of the FS dataset

Attribute	DM	Ave	Dis	RV
Protein	**P30**	0.205	0.240	11
Enzyme	**P30**	0.150	0.080	11
Decomposition Enzyme	N	N	N	N
Synthesis enzyme	**P30**	0.080	0.060	11
The post hydrogen enzyme	B62	0.630	N	11
Phosphorylation enzyme	B62	0.734	0.240	12
Transcriptase enzyme	NA	NA	NA	NA
Catalyze enzyme	NA	NA	NA	NA
Intravital material	**P30**	0.170	0.130	11
Molecule	N	N	N	N
Compound	N	N	N	N
Bound region	N	N	N	N
Component of medicine	N	N	N	N
Virus	B45,B60 B90	0.120	0.170	11
Amino acid	N	N	N	N

Table 7: Results of using the FD dataset

Attribute	DM	Ave	Dis	RV
Protein	**P70**	0.290	0.270	11
Enzyme	B80	0.330	0.180	11
Decomposition Enzyme	B45	0.660	0.140	12
Synthesis enzyme	B80	0.150	0.110	11
The post hydrogen enzyme	B80	0.420	0.230	11
Phosphorylation enzyme	N	N	N	N
Transcriptase enzyme	N	N	N	N
Catalyze enzyme	B45,B60	0.670	0.140	40
Intravital material	B80	0.360	0.170	11
Molecule	B90	0.350	0.170	11
Compound	B90	0.215	0.060	11
Bound region	N	N	N	N
Component of medicine	B90	0.420	0.400	11
Virus	B45,B60	0.120	0.150	11
Amino acid	N	N	N	N

On comparing the results of the two datasets, each distance matrix with a high score was defined as optimal for each attribute. In the blank representation denoted by “N” and “NA”, since the average score is low, the attribute with an undetermined optimum distance matrix was excluded (Table 8).

Table 8: Results of using the Distance Matrix

Attribute	Optimal DM
Protein	B80
Enzyme	B80
Decomposition Enzyme	B80
Synthesis enzyme	B80
The post hydrogen enzyme	B50
Phosphorylation enzyme	B80
Transcriptase enzyme	B90
Catalyze enzyme	B45,B60
Component of medicine	B45
Intravital material	P250
Virus	B50

BLOSUM 80 was selected as optimal for the many types of enzyme sequence. BLOSUM 45 and BLOSUM 50 are suitable not only enzyme, but also other types such as medicine and virus.

6. Conclusion

To identify the optimal distance matrix, we performed a comparative analysis by standardizing the alignment algorithm. Six BLOSUM and 3 PAM distance matrices were utilized with the BAliBASE 3.0 as the benchmark database for experimental data. Our comparative analysis revealed each optimal distance matrix for each biological characteristic. The present results indicate that there are robust distant matrices applicable to any attribute of amino acid sequences and are limited in accordance with the structure of the sequence.

Acknowledgments

We would like to thank Editage (www.editage.jp) for English language editing.

References

[1] Hirosawa M, Totoki Y, Hoshida M, Ishikawa M, “Comprehensive study on iterative algorithms of multiple sequence alignment,” CABIOS, vol.11 No.1, pp.13-18, 1995.

[2] Wallace I.M., O, Sullivan O. and Higgins D.G, “Evaluation of iterative alignment algorithms for multiple alignment,” Bioinformatics, vol.21, No.8, pp.1408-1414, 2005.

[3] Mizuguchi K., “HOMSTRAD: a database of protein structure alignments for homologous families,” Protein Science, vol.7, pp.2469-2471, 1998.

[4] Wakatsu D, Okazaki T, “Statistical Comparative Study of Multiple Sequence Alignment Scores of Iterative Refinement Algorithms,” IPSJ Transactions on Bioinformatics, vol. 2, 74–82, 2009.

[5] Smith, T. F. and Waterman, M. S, “Identification of Common Molecular Subsequences,” Journal of Molecular Biology, vol. 147, No. 1, pp.195-197, 1981.

[6] Needleman, Saul B. & Wunsch, Christian D, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” Journal of Molecular Biology, vol.48, No.3, pp.443-453, 1970.

[7] S Henikoff, J G Henikoff, “Amino acid substitution matrices from protein blocks,” PNAS, vol.89, No.22, pp.10915-10919, 1992.

[8] BAliBASE http://www.lbgi.fr/balibase/ (accessed 2008)

Ayako OHSHIRO received the B.S. and M.S. degrees from University of the Ryukyus in 2009 and 2011, respectively. She took Ph.D. from University of Ryukyus in 2017 and she has been assistant professor at University of the Ryukyus. Her research interests are analysis, modeling, and comparative study of data using information statistics.

Takeo OKAZAKI took his B.Sc. and M.Sc. degrees from Kyusyu University in 1987 and 1989, respectively. He took Ph.D. from University of the Ryukyus in 2014. He has been a professor at University of the Ryukyus. His research interests are statistical data normalization for analysis, and casual relationship analysis.