Hindawi
BioMed Research International
Volume 2018, Article ID 1718046, 9 pages
https://doi.org/10.1155/2018/1718046

Research Article

Disease Sequences High-Accuracy Alignment Based on the Precision Medicine

ManZhi Li,[1] HaiXia Long,[2] HongTao Wang,[1] HaiYan Fu,[2] Dong Xu,[2] YouJian Shen,[1] YuHua Yao,[1] and Bo Liao[1]

[1]*School of Mathematics and Statistics, Hainan Normal University, Haikou, Hainan 571158, China*
[2]*School of Information Science Technology, Hainan Normal University, Haikou, Hainan 571158, China*

Correspondence should be addressed to HaiXia Long; 64169486@qq.com

Received 22 November 2017; Accepted 18 January 2018; Published 22 February 2018

Academic Editor: Tao Huang

High-accuracy alignment of sequences with disease information contributes to disease treatment and prevention. The results of multiple sequence alignment depend on the parameters of the objective function, including gap open penalties (GOP), gap extension penalties (GEP), and substitution matrix (SM). Firstly, the theory parameter formulas relating to GOP, GAP, and SM are inferred, combining unaligned sequence length, number, and identity. Secondly, we tested the rationality of the theory parameter formulas, with experiment on the ClustalW and MAFFT program. In addition, we obtained a group of MAFFT program parameters according to the formulas proposed. The results of all experiments show that the SPS (sum-of-pair score) obtained from theory parameters is better than the SPS obtained from the default parameters of ClustalW and MAFFT. In both theory and practice, our method to determine the parameters is feasible and efficient. These can provide high-accuracy alignment results for precision medicine.

1. Introduction

In 2015, US President Barack Obama stated his intention to fund a United States national "Precision Medicine Initiative" [1, 2]. A short-term goal of the Precision Medicine Initiative is to expand cancer genomics to develop better prevention and treatment methods. With the explosive growth of medical data, the complexity of disease, and the demand of personalized medicine, the research results of genome sequencing are changing the process of disease treatment. Multiple sequence alignment (MSA) is more and more important.

Multiple sequence alignment (MSA) has wide applications in sequence analysis, gene recognition, protein structure prediction, and reconstructing the phylogenetic tree [3]. Notredame [4] stated that the most modern programs for constructing MSA consist of two components: (1) an objective function to assess the quality of candidate alignment and (2) an optimization procedure for identifying the highest scoring alignment with respect to the chosen objective function. Currently, MSA has three main objective functions: (1) the sum-of-pairs score function (SPS), (2) the consensus function, and (3) the tree function. The SPS function is the most commonly used objective function, and its parameters include substitution matrix and gap opening penalties (GOP) and gap extending penalties (GEP).

The parameters of the objective function have generated many discussions on how to obtain optimal parameters. Thompson et al. [5] determined that substitution matrices vary at different alignment stages according to the divergence of sequences to be aligned. Residue-specific gap penalties and gap penalties in hydrophilic regions, which have been locally reduced, can cause new gaps to appear in potential loop regions rather than in a regular secondary structure. Reese and Pearson [6] discussed the relational formula between the PAM distance and PAM matrix as well as the gap penalty. Madhusudhan et al. [7] proposed the variable penalty formula according the structure of sequence based on dynamic programming. However, these formulas are not widely used. Gondro and Kinghorn [8] indicated that gap penalty parameters were determined by experience. At present, it is no theoretical framework to determine the optimum parameters. The current parameters pertaining

to the objective function in most literature are empirical values which are independently associated with the sequences [9]. BALiBASE is a database of manually refined multiple sequence alignments [10] and is usually used to test performance of MSA method [11].

Many open source online alignment tools are available that can align hundreds of thousands of sequences in hours. These include CLUSTAL Omega, T-COFFEE, and MAFFT, [5, 12–14] and often become the primary source of sequence alignment solution. However, these MSA tool results strongly depend on the gap penalty and substitution matrix. Different parameter combinations can obtain different MSA results. The majority of users use a single default parameter when applying these alignment tools, but the results are not the best. Moreover, an effective methodology has not yet been developed to directly determine an MSA optimal parameter, which means current online tools cannot guarantee the best solution. However, when compared with other MSA alignment tools, MAFFT has the advantage of simple input parameters and obtains better results than the other tools [12, 13]. This paper uses MAFFT as the basic experimental tool to verify the accuracy of the original formulas presented herein as they relate to the substitution matrix and the gap penalty.

2. Sum-of-Pairs (SP) Objective Function

The sum-of-pairs (SP) function is commonly used as an objective function for MSA and is derived as

$$\text{score} = \sum \text{Residue} - \sum \text{penalty}, \tag{1}$$

where the score is >0. When the score is higher, the accuracy of MSA is higher [15]. $\sum$ Residue > 0 represents the total score of amino acid residues in the alignment sequence. $\sum$ penalty is the total penalty score due to inserting gap and $\sum$ penalty $>$ 0.

$\sum$ Residue is calculated as

$$\sum \text{Residue} = \sum_{h=1}^{L} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{Cost}\left(S_i, S_j\right), \tag{2}$$

where S_{ih} is the h residue of the i sequence, L is the length of the aligned sequences, and k is the number of the sequences.

$$\text{Cost}\left(S_i, S_j\right) = \begin{cases} S_{aa} = \text{Score}\,(a, a) & \text{if } a = a \text{ (residues are matched)} \\ S_{ab} = \text{Score}\,(a, b) & \text{if } a \neq \text{“}-\text{”}, \; b \neq \text{“}-\text{”} \text{ (residues are mismatched)} \\ S_{a-} = \text{Score}\,(a, -) = 0 & \text{if } a \neq \text{“}-\text{”} \left(\text{residue and gap}\right). \end{cases} \tag{3}$$

Cost is computed by a substitution matrix. Currently, two main kinds of substitution matrices are available: PAM and BLOSUM. The BLOSUM series applies to this research. In substitution matrices, S_{aa} are different from each other. When the residues are mismatched, S_{ab} are also different from each other. But, in the process of simplifying the calculation, we need to use a precise and representative numerical value to represent the characteristics of the matrix. The average value can be a good characteristic representing a group of different data. Therefore, using the average value mean(S_{aa}) of S_{aa} represents the match of the matrix and using an average value mean(S_{ab}) of S_{ab} represents the mismatch of the matrix.

The calculation of $\sum$ penalty is divided into two categories: linear penalty and affine penalty. Linear penalty penalizes the same score for each gap. Affine penalty is commonly used because it is biologically meaningful [16–18]. The gap is divided into two types: gap open penalty (GOP) and gap extension penalty (GEP), so the affine penalty formula is given as

$$\sum \text{penalty} = N_{\text{GOP}} \cdot \text{GOP} + N_{\text{GEP}} \cdot \text{GEP}, \tag{4}$$

where N_{GOP} is the number of GOP, N_{GEP} is the number of GEP, and GOP > GEP.

3. The Theory Parameters Determination of SP Function for MSA

Symbol Description. The number of unaligned sequences is m. The length of the longest sequence is $\text{len}_{\max}$. The length of the shortest sequence is $\text{len}_{\min}$. The mean identity is iden. The number of amino acid residues matched is $\text{num}_{\text{match}} = (m(m-1)/2) \cdot \text{len}_{\min} \cdot \text{iden}$. After alignment, the number of gaps inserted into each sequence is num_{gap}.

Table 1 summarizes the ratio of the longest sequence and the number of gaps inserted into the sequence of each data set in BAliBASE 2.0 and BAliBASE 3.0. It shows that the number of gaps in the longest sequence is not more than 0.2 times the length of the longest sequence. That is, the number of gaps in each sequence is $\text{num}_{\text{gap}} \leq \text{int}\,(0.2 \cdot \text{len}_{\max}) + \text{len}_{\max} - \text{len}_{\min}$, and int is the rounding function. Figure 1 shows how the sequence length and the number of gaps num_{gap} are related.

Figure 1 is an example. If $\text{len}_{\text{align}} = 25$, $\text{len}_{\max} = 21$, and $\text{len}_{\min} = 7$, the number of gaps inserted into the longest sequence is $\text{num}_{\text{gap}} = \text{len}_{\text{align}} - \text{len}_{\max} = 25 - 21 = 4$, and the ratio between the sequence and gaps is ratio $= (\text{len}_{\text{align}} - \text{len}_{\max})/\text{len}_{\max} = 4/21 = 0.19$. The number of gaps in the sequence is $\text{num}_{\text{gap}} \leq \text{int}\,[0.2 \cdot \text{len}_{\max}]$. The number of gaps inserting the shortest sequence is $\text{num}_{\text{gap}} = \text{len}_{\text{align}} - \text{len}_{\min} = 25 - 7 = 18$, and the number of gaps in sequence is $\text{num}_{\text{gap}} \leq$

Table 1: Ratio of the longest sequence and the number of gaps inserted into the sequence.

	BALIBASE 2.0					BALIBASE 3.0	
Data set	Test 1	Test 2	Test 3	Ref 2	Ref 3	RV11	RV12
Mean (ratio)	0.0769	0.0764	0.0744	0.1439	0.1612	0.1938	0.0784

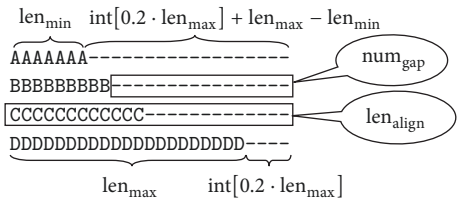


Figure 1: The relationship between the sequence length and the number of gaps.

$\text{int}[0.2 \cdot \text{len}_{\max}] + \text{len}_{\max} - \text{len}_{\min}$. The number of gaps in other sequences is $\text{num}_{\text{gap}} \le \text{int}[0.2 \cdot \text{len}_{\max}] + \text{len}_{\max} - \text{len}_{\min}$.

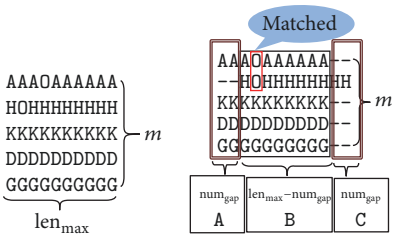
The following parameter formulas are inferred according to information obtained from Figure 2. Figure 2(a) has the best state unaligned sequence. Each sequence has the same length and no gaps. The longest length of any unaligned sequence is 10, so the number of gaps inserted can go up to 2. Figure 2(b) shows the worst alignment results (inserting maximum gap and minimum matching). If the score of Figure 2(b) is higher than the score of Figure 2(a), the parameters of the objective function meet all cases of alignment, because the situation in Figure 2 is the worst alignment.

3.1. Substitution Matrix Theory Formula. According to (1), the SP score of unaligned sequences is

$$\begin{aligned} \text{score}_{\text{begin}} &= \sum \text{Residue} - \sum \text{penalty} \\ &= \frac{m(m-1)}{2} \cdot \text{len}_{\max} \cdot S_{ab} \end{aligned} \tag{5}$$

and according to (1) and Figure 2(b), the following equations can be obtained:

$$\begin{aligned} &\text{score}_{\text{end}} = \text{score } A + \text{score } B + \text{score } C, \\ &\text{score } A = \alpha \cdot \frac{(m-1)(m-2)}{2} \cdot \text{num}_{\text{gap}} \cdot S_{ab}, \\ &\text{score } B \\ &\quad = \left[\frac{m(m-1)}{2} \left(\text{len}_{\max} - \text{num}_{\text{gap}} \right) - \text{num}_{\text{match}} \right] \\ &\qquad \cdot S_{ab} + \beta \cdot \text{num}_{\text{match}} \cdot S_{aa}, \\ &\text{score } C = \sum \text{penalty}. \end{aligned} \tag{6}$$



(a) The best state unaligned sequences (b) The worst alignment results

Figure 2: Unalignment and alignment.

So, the SP score of the aligned sequences is

$$\begin{aligned} \text{score}_{\text{end}} = &\left[\frac{m(m-1)}{2} \cdot \left(\text{len}_{\max} - \text{num}_{\text{gap}} \right) \right. \\ &\left. - \text{num}_{\text{match}} + \alpha \cdot \frac{(m-1)(m-2)}{2} \cdot \text{num}_{\text{gap}} \right] \cdot S_{ab} \\ &+ \beta \cdot \text{num}_{\text{match}} \cdot S_{aa} - \sum \text{penalty}. \end{aligned} \tag{7}$$

In theory, the alignment score must be greater than the unaligned sequence score,

$$\text{score}_{\text{begin}} \le \text{score}_{\text{end}}. \tag{8}$$

That is,

$$\begin{aligned} &\frac{m(m-1)}{2} \cdot \text{len}_{\max} \cdot S_{ab} \le \left[\frac{m(m-1)}{2} \right. \\ &\quad \cdot \left(\text{len}_{\max} - \text{num}_{\text{gap}} \right) - \text{num}_{\text{match}} + \alpha \\ &\quad \left. \cdot \frac{(m-1)(m-2)}{2} \cdot \text{num}_{\text{gap}} \right] \cdot S_{ab} + \beta \cdot \text{num}_{\text{match}} \\ &\quad \cdot S_{aa} - \sum \text{penalty}. \end{aligned} \tag{9}$$

Equation (9) can be simplified as

$$\begin{aligned} S_{aa} \ge &\left[\frac{(\alpha m - 2\alpha - m)(1-m)}{2\beta} \cdot \frac{\text{num}_{\text{gap}}}{\text{num}_{\text{match}}} + \frac{1}{\beta} \right] \\ &\cdot S_{ab}. \end{aligned} \tag{10}$$

The formula of the substitution matrix is shown in (10), which can be simplified as

$$\text{reference} \ge \text{calc}. \tag{11}$$

The rationality of the substitution matrix can be judged according to (11).

3.2. GOP and GEP Theory Formulas. Based on the affine penalty, num_{gap} is the number of gaps of each sequence; let us suppose that the number of gaps in each sequence is λ times as the number of GOP, so $N_{\text{GOP}} = m \cdot (1/\lambda) \cdot \text{num}_{\text{gap}}$ and

$N_{\text{GEP}} = m \cdot (1 - 1/\lambda) \cdot \text{num}_{\text{gap}}$. Because GOP > GEP, we accept that GOP $= n \cdot$ GEP, where λ, n is the positive integer, so

$$\begin{aligned} \sum \text{penalty} &= N_{\text{GOP}} \cdot \text{GOP} + N_{\text{GEP}} \cdot \text{GEP} \\ &= \frac{n + \lambda - 1}{n\lambda} \cdot m \cdot \text{num}_{\text{gap}} \cdot \text{GOP}. \end{aligned} \tag{12}$$

According to (12), (9) can be expressed as follows:

$$\begin{aligned} &\frac{(\alpha m - 2\alpha - m)(m-1)}{2} \cdot \text{num}_{\text{gap}} \cdot S_{ab} + \text{num}_{\text{match}} \\ &\cdot (\beta S_{aa} - S_{ab}) \geq \sum \text{penalty} \Longrightarrow \\ &\text{GOP} \leq \left[\frac{(\alpha m - 2\alpha - m)(m-1)}{2} \text{num}_{\text{gap}} S_{ab} \right. \\ &\left. + \text{num}_{\text{match}} (\beta S_{aa} - S_{ab}) \right] \\ &\cdot \frac{n\lambda}{m(n + \lambda - 1) \cdot \text{num}_{\text{gap}}}. \end{aligned} \tag{13}$$

Equation (13) is the upper limit of GOP and the lower limit is GOP > 0.

If the upper limit of GOP is multiplied by weight coefficient ω and $0 < \omega < 1$, the estimation formula of GOP is

$$\begin{aligned} \text{GOP} = \omega \cdot &\left[\frac{(\alpha m - 2\alpha - m)(m-1)}{2} \text{num}_{\text{gap}} S_{ab} \right. \\ &\left. + \text{num}_{\text{match}} (\beta S_{aa} - S_{ab}) \right] \\ &\cdot \frac{n\lambda}{m(n + \lambda - 1) \cdot \text{num}_{\text{gap}}}, \end{aligned} \tag{14}$$

where $\text{num}_{\text{match}} = (m(m-1)/2) \cdot \text{len}_{\text{min}} \cdot \text{iden}$, $\text{num}_{\text{gap}} = \text{int}(0.2 \cdot \text{len}_{\text{max}}) + \text{len}_{\text{max}} - \text{len}_{\text{min}}$, and int is a rounding function. len_{min} is the length of the shortest sequence in the unaligned sets, and iden is the mean identity of unaligned sets.

The estimation formula of GEP is

$$\text{GEP} = \frac{\text{GOP}}{n}. \tag{15}$$

The optimal value of each weight coefficients λ, n, ω, α, and β in (14) and (15) can be obtained through the following experiments.

4. Simulation and Results

In order to test the rationality of the parameter formulas and determine the optimal value of each weight coefficient, we designed the following experiments on the BAliBASE 2.0 and BAliBASE 3.0.

4.1. Experiment Setting. BAliBASE version 2.0 [10] is an improved version, extended from version 1 with 167 reference alignments to over 2100 sequences, which also features eight reference sets. Because all the reference alignments of BAliBASE are aligned by the manual, it often used to test algorithms [19–21]. Because our study is based on the global SP function, in this article, we used 113 reference alignments in References 1–3 as test objects. BAliBASE version 3.0 has the most widely used multiple alignment benchmark. The database contains 218 multiple protein sequence alignments, which have been divided into five reference sets. The first reference set includes equidistant sequences, whose identity is less than 20% (RV11) or between 20 and 40% (RV12) [22]. Other references have no similarity information. Because the formulas proposed in this paper need similarity of sequences, BAliBASE 2.0 and BAliBASE 3.0 (RV11 and RV12) were both used to establish data sets.

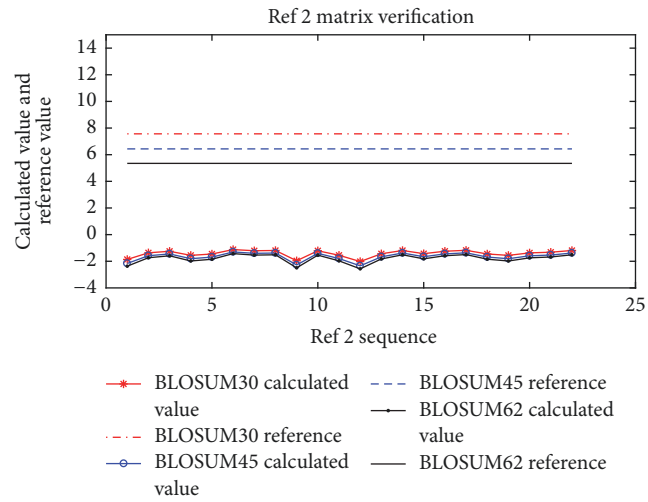


Figure 3: The results of the verification of substitution matrix (11).

SPS (sum-of-pair score) works as an objective function, which can determine score increases if sequences are correctly aligned. If the SPS is higher, the results of alignment are close to the reference alignment and can be even better than the reference alignment [20]. To test the rationality of presented formulas and to determine the optimal parameters combination of MSA tools, the most popular alignment program, MAFFT [16], is used in this research. The alignment results are obtained through the Perl programming language. The MAFFT program has some advantages: (1) the number of MAFFT program parameters is less and is easy to control, using only substitution matrices, GOP and GEP, (2) through Perl, the MAFFT program can batch align, and (3) alignment accuracy is for the most part better than CW, MUSCULE, and TCOFFEE.

In our experiment, $1 \leq \text{GOP} \leq 20, 0 \leq \text{GEP} \leq \text{GOP}/2$. The GOP step is 1, the GEP step is 0.2, and the substitution matrices are BLOSUM30, BLOSUM45, and BLOSUM62. For each group of sequences, through batch processing, the number of alignment results is 1,590 because there are 1,590 different combined parameter patterns.

4.2. Experiment Results

4.2.1. The Verification of Substitution Matrix Formula. This section shows how the rationality of the substitution matrix was established (see (11)). Figure 3 illustrates the calculated

Table 2: The number of sequences meeting the substitution matrix requirements (see (11)).

	Sequence number	Reference alignment number	BLOSUM30 qualified number (rate)	BLOSUM45 qualified number (rate)	BLOSUM62 qualified number (rate)
Reference 1	4-5	78	78 (100%)	78 (100%)	78 (100%)
Reference 2	14–19	22	22 (100%)	22 (100%)	22 (100%)
Reference 3	> 20	12	12 (100%)	12 (100%)	12 (100%)

Table 3: Determination of the value of $n, \lambda, \omega, \alpha, \beta$.

$n = 5, \lambda = 3, \alpha = 0.2, \beta = 0.9$	BLOSUM30		BLOSUM45		BLOSUM62	
ω	num	SPS	num	SPS	num	SPS
0.01	7	0.7586	11	0.7768	9	0.7697
0.02	9	0.761	10	0.7769	8	0.7692
0.03	10	0.7643	11	0.7795	10	0.7703
0.04	12	0.782	15	0.7886	12	0.7745
0.05	14	0.7843	**19**	**0.8003**	15	0.7846
0.06	14	0.7805	17	0.7924	17	0.7864
0.07	14	0.7767	16	0.7896	16	0.786
0.08	14	0.7728	16	0.784	14	0.7821
0.09	14	0.7668	15	0.7804	14	0.7777
0.1	14	0.764	15	0.78	15	0.7826
0.01	7	0.7586	11	0.7768	9	0.7697
0.02	9	0.7614	10	0.7769	9	0.77
0.03	10	0.7681	12	0.7818	12	0.7744
0.04	15	0.7874	15	0.7877	13	0.7858
0.05	13	0.7783	15	0.79	16	0.7918
0.06	13	0.7736	14	0.7845	14	0.7859
0.07	13	0.7732	14	0.7781	12	0.7819
0.08	13	0.7709	16	0.7846	12	0.7713
0.09	14	0.7779	15	0.7781	13	0.7751
0.1	14	0.7731	15	0.7795	14	0.7779

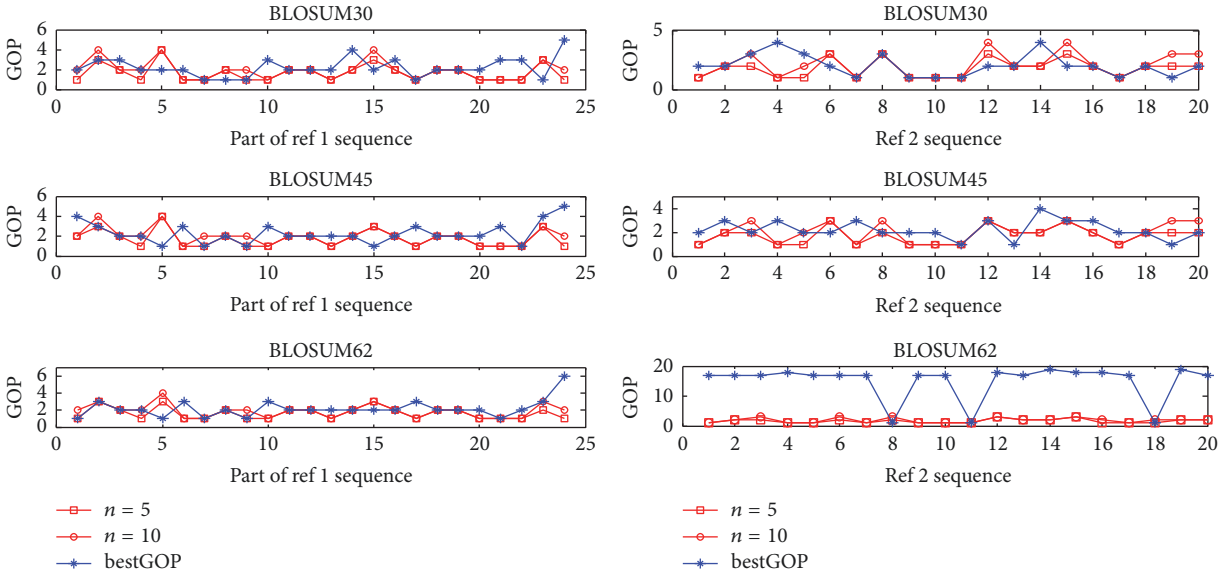
value and reference value of each of the three substitution matrices for Reference 2 (note: the other figures are similar to Figure 3). According to (11), when the reference value is greater than the reference value, the substitution matrix is rationality. It is shown that BLOSUM30, BLOSUM45, and BLOSUM62 meet the requirements of all sequences.

Table 2 lists the number of sequences meeting the substitution matrix sequence requirements (see (11)). It is shown that three BLOSUM substitution matrices meet all the sequences for References 1–3.

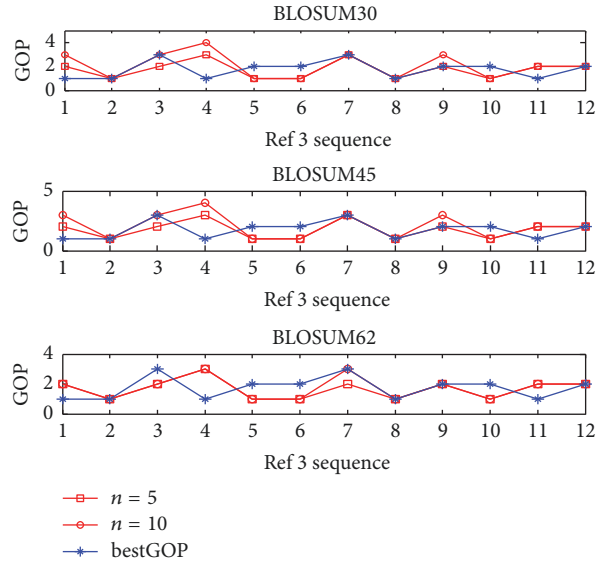
4.2.2. The Verification of Gap Penalty Formulas. Based on the SPS and MAFFT program (MAFFT-7.220-WIN64 version), we tested the rationality of (14) and (15). The optimum of GOP corresponded to the maximal SPS illustrated in Figure 4. From Figure 4, we can conclude the following: the GOP theory values inferred from (14) and (15) almost coincide with the optimal of GOP, so (14) can calculate the optimal value of GOP.

Table 3 statistics show the number of sequences in Reference 1 (Test 2), which meet the theory parameter requirements corresponding to SPS, which are greater than the default parameters corresponding to SPS. In Test 2, there are 24 sequences. Table 3 shows that when $\lambda = 3$, $\alpha = 0.2$, $\beta = 0.9$, and $n = 5$, the number of sequences is greater than $\lambda = 3, \alpha = 0.2, \beta = 0.9$, and $n = 10$. The best result is indicated in Blosum45, num 19, with an SPS of 0.8003 (in Table 3 set in bold face font). For Test 2 sequence sets, $\lambda = 3$, $n = 5$ is relatively rational and corresponds to $\omega = 0.05$. The other sequence sets can also obtain the value of n, λ, ω, α, and β, which are listed in Table 4.

4.2.3. Finding Optimal Value of Other Parameters in Derivation Formula. From the aforementioned experiments, we can determine the substitution matrix and n, λ, and ω in (14). The other parameters are related to the sequences where λ is the ratio of GOP and num_{gap}, and $\text{num}_{\text{gap}} = \text{int}(0.2 \cdot \text{len}_{\max}) + \text{len}_{\max} - \text{len}_{\min}$. The number of GOP is limited and it will



(a) The results of verification of GOP/GEP in Reference 1 sequences (b) The results of verification of GOP/GEP in Reference 2 sequences

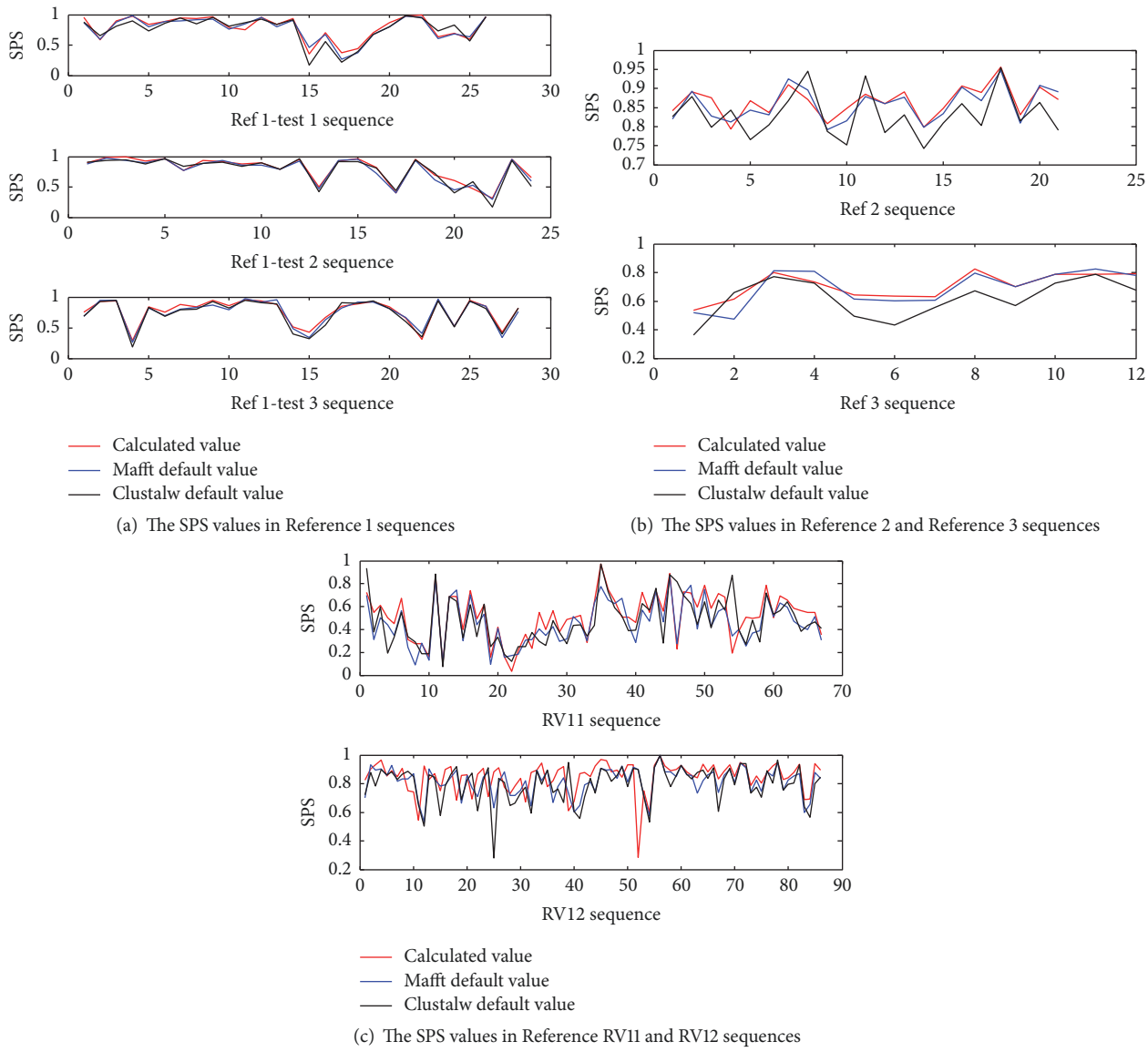


(c) The results of verification of GOP/GEP in Reference 3 sequences

Figure 4: The results of verification of GOP/GEP in (14) and (15).

Table 4: Optimal GOP/GEP/matrix.

Sequence set	Ref 1-test 1	Ref 1-test 2	Ref 1-test 3	Ref 2	Ref 3
Sequence row	4-5	4-5	4-5	14–19	>20
Sequence length (bp)	<100	100–300	>300	50–600	60–600
ω	0.03	0.05	0.08	0.02	0.02
n	5	5	10	10	10
Matrix	BLOSUM45	BLOSUM45	BLOSUM62	BLOSUM45	BLOSUM45
λ			3		
α			0.2		
β			0.9		



(a) The SPS values in Reference 1 sequences

(b) The SPS values in Reference 2 and Reference 3 sequences

(c) The SPS values in Reference RV11 and RV12 sequences

Figure 5: The SPS values are from MAFFT theory parameters, MAFFT default parameters, and CLUSTALW default parameter.

not increase too much, while the distribution of GEP is more concentrated. These parameters are more consistent with the biological characteristics of multiple sequence alignment.

Optimal parameters and the SPS value are listed in Table 4. The optimal value of weight coefficient in our proposed formula is located in Table 4. Using a weight coefficient, we can obtain the optimal of GOP, GEP, and MATRIX parameters. The number of sequences corresponding to SPS is also listed in Table 4.

Figure 5 shows that, for each SPS value sequence obtained from theory parameters, we inferred default parameters of MAFFT (MAFFT-7.220-WIN64 version) and CLUSTALW (CLUSTALW-2.1-WIN version). The SPS obtained by the MAFFT program are better than the CLUSTALW program on the default parameters. So we chose the MAFFT program as our test method. The SPS obtained by our theory parameters were better than the default parameters of MAFFT and CLUSTALW. Thus, the theory parameters we propose can optimize the results of MSA.

Table 5 shows the SPS mean values of References 1–3 sequences of BAliBASE 2.0 and RV11/RV12 of BAliBASE 3.0. The alignment sequences obtained from MAFFT default parameters, CLUSTALW default parameters, and MAFFT theory parameters are those proposed in this study. It is shown that SPS values obtained by MAFFT default parameters are better than SPS values obtained by CLUSTALW default parameters. The SPS values obtained using our theory parameters are the best. So, the theory parameters optimized the results of MSA.

5. Conclusions

This paper clearly shows that the parameters of MSA tools influence MSA results. These parameters not only include

Table 5: SPS mean value.

	BaliBASE 2.0					BaliBASE 3.0	
Data set	Ref 1 (test 1)	Ref 1 (test 2)	Ref 1 (test 3)	Ref 2	Ref 3	RV11	RV12
MAFFT default parameters	0.7749	0.7743	0.7460	0.8584	0.6938	0.4582	0.8142
CW default parameters	0.7614	0.7732	0.7340	0.8311	0.6189	0.4758	0.7966
MAFFT theory parameters	0.7918	0.8003	0.7652	0.8655	0.7073	0.5183	0.8449

substitution matrices, GOP, and GEP but also include the length, number, and identity of sequences. Our goal was to find a group of combined optimal parameters. Based on the SP function, we established a series of formulas which can determine the value of substitution, GOP, and GEP. In order to test the rationality of the formulas, our experiments were conducted in the MAFFT program base or in the BAliBASE 2.0 and BAliBASE 3.0 (RV11 and RV12) database. Moreover, we obtained the optimal value of the substitution matrices, GOP and GEP, and these values proved to be better than the default values of the MAFFT program. After the theory analysis and experimental analysis, we can conclude that the proposed method can effectively solve the MSA parameter problems and improve MSA accuracy, which can provide more accuracy information for precision medicine in disease analysis and prediction.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

ManZhi Li and HaiXia Long contributed equally to this work. ManZhi Li and HaiXia Long carried out the multiple sequence alignment parameters studies, participated in the experiments, and drafted the manuscript; these authors contributed equally to this work. HaiYan Fu and HongTao Wang participated in the design of the study and performed the statistical analysis. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the China Scholarship Council, the National Natural Science Foundation of China (no. 61762034, no. 71461008, no. 61663007, and no. 61163042), and the HaiNan Province Natural Science Foundation (no. 614235, no. 617122, and no. 20166222).

References

[1] T. P. Conrads and E. F. Petricoin, "The Obama Administration's Cancer Moonshot: A Call for Proteomics," *Clinical Cancer Research*, vol. 22, no. 18, pp. 4556–4558, 2016.

[2] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *The New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[3] Q. Le, F. Sievers, and D. G. Higgins, "Protein multiple sequence alignment benchmarking through secondary structure prediction," *Bioinformatics*, vol. 33, no. 9, pp. 1331–1337, 2017.

[4] C. Notredame, "Recent progress in multiple sequence alignment: A survey," *Pharmacogenomics*, vol. 3, no. 1, pp. 131–144, 2002.

[5] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[6] J. T. Reese and W. R. Pearson, "Empirical determination of effective gap penalties for sequence comparison," *Bioinformatics*, vol. 18, no. 11, pp. 1500–1507, 2002.

[7] M. S. Madhusudhan, M. A. Marti-Renom, R. Sanchez, and A. Sali, "Variable gap penalty for protein sequence-structure alignment," *Protein Engineering, Design and Selection*, vol. 19, no. 3, pp. 129–133, 2006.

[8] C. Gondro and B. P. Kinghorn, "A simple genetic algorithm for multiple sequence alignment," *Genetics and Molecular Research*, vol. 6, no. 4, pp. 964–982, 2007.

[9] D. DeBlasio and J. Kececioglu, "Parameter advising for multiple sequence alignment," *BMC Bioinformatics*, vol. 16, no. Suppl 2, p. A3, 2015.

[10] J. D. Thompson, F. Plewniak, and O. Poch, "BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs," *Bioinformatics*, vol. 15, no. 1, pp. 87-88, 1999.

[11] R. C. Edgar, "Quality measures for protein alignment benchmarks," *Nucleic Acids Research*, vol. 38, no. 7, Article ID gkp1196, pp. 2145–2153, 2010.

[12] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[13] K. Katoh and H. Toh, "Recent developments in the MAFFT multiple sequence alignment program," *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 286–298, 2008.

[14] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

[15] K. Reinert, J. Stoye, and T. Will, "An iterative method for faster sum-of-pairs multiple sequence alignment," *Bioinformatics*, vol. 16, no. 9, pp. 808–814, 2000.

[16] O. Gotoh, "Multiple sequence alignment: Algorithms and applications," *Advances in Biophysics*, vol. 36, pp. 159–206, 1999.

[17] M. Kaya, A. Sarhan, and R. Alhajj, "Multiple sequence alignment with affine gap by using multi-objective genetic algorithm," *Computer Methods and Programs in Biomedicine*, vol. 114, no. 1, pp. 38–49, 2014.

[18] Q. Zou, X. Shan, and Y. Jiang, "A Novel Center Star Multiple Sequence Alignment Algorithm Based on Affine Gap Penalty and K-Band," *Physics Procedia*, vol. 33, pp. 322–327, 2012.

[19] A. Bahr, J. D. Thompson, J.-C. Thierry, and O. Poch, "BAliBASE (Benchmark Alignment dataBASE): Enhancements for repeats,

transmembrane sequences and circular permutations," *Nucleic Acids Research*, vol. 29, no. 1, pp. 323–326, 2001.

[20] F. M. Ortuño, O. Valenzuela, F. Rojas et al., "Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: Structural information, non-gaps percentage and totally conserved columns," *Bioinformatics*, vol. 29, no. 17, pp. 2112–2121, 2013.

[21] F. Naznin, R. Sarker, and D. Essam, "Vertical decomposition with Genetic Algorithm for Multiple Sequence Alignment," *BMC Bioinformatics*, vol. 12, article no. 353, 2011.

[22] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, "BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark," *Proteins: Structure, Function, and Genetics*, vol. 61, no. 1, pp. 127–136, 2005.