Computational Biology and Chemistry 35 (2011) 308–318

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Comparison of linear gap penalties and profile-based variable gap penalties in profile–profile alignments

Chuan Wang[a,b], Ren-Xiang Yan[a,b], Xiao-Feng Wang[a,b], Jing-Na Si[a,b], Ziding Zhang[a,b,*]

[a] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China
[b] Bioinformatics Center, College of Biological Sciences, China Agricultural University, Beijing 100193, China

ARTICLE INFO

Article history:
Received 5 February 2011
Received in revised form 6 May 2011
Accepted 11 July 2011

Keywords:
Gap distribution
Indel frequency profile
Parameter optimization

ABSTRACT

Profile–profile alignment algorithms have proven powerful for recognizing remote homologs and generating alignments by effectively integrating sequence evolutionary information into scoring functions. In comparison to scoring function, the development of gap penalty functions has rarely been addressed in profile–profile alignment algorithms. Although indel frequency profiles have been used to construct profile-based variable gap penalties in some profile–profile alignment algorithms, there is still no fair comparison between variable gap penalties and traditional linear gap penalties to quantify the improvement of alignment accuracy. We compared two linear gap penalty functions, the traditional affine gap penalty (AGP) and the bilinear gap penalty (BGP), with two profile-based variable gap penalty functions, the Profile-based Gap Penalty used in SP^5 (SPGP) and a new Weighted Profile-based Gap Penalty (WPGP) developed by us, on some well-established benchmark datasets. Our results show that profile-based variable gap penalties get limited improvements than linear gap penalties, whether incorporated with secondary structure information or not. Secondary structure information appears less powerful to be incorporated into gap penalties than into scoring functions. Analysis of gap length distributions indicates that gap penalties could stably maintain corresponding distributions of gap lengths in their alignments, but the distribution difference from reference alignments does not reflect the performance of gap penalties. There is useful information in indel frequency profiles, but it is still not good enough for improving alignment accuracy when used in profile-based variable gap penalties. All of the methods tested in this work are freely accessible at http://protein.cau.edu.cn/gppat/.

1. Introduction

Profile–profile alignment algorithms are powerful tools for sensitively detecting protein remote homology relationships with improved alignment accuracy (Ohlson et al., 2004). Traditional protein sequence alignment methods use substitution matrices to measure the similarity of amino acid pairs, while profile–profile alignment methods require a profile-based scoring function to measure the similarity of profile vector pairs. Both types of alignment methods employ gap penalty functions. Many different profile–profile alignment scoring functions have been developed and evaluated, yielding significant improvements for both alignment accuracy and fold recognition (Edgar and Sjolander, 2004; Ohlson et al., 2004; Wang and Dunbrack, 2004). Some of these scoring functions are integrated with structural information, such as secondary structure, solvent accessibility, backbone dihedral torsion angles, hydrophobic indices, and structural profiles of templates, either predicted from the query sequence or obtained from the template structure. With the integration of structural information, profile–profile alignment has become a powerful approach for recognizing distant homologs and providing high-quality alignments for the prediction of protein structures as a single method (not meta-server), e.g., the SPARKS and SP series developed by Zhou and co-workers (Liu et al., 2007; Zhang et al., 2008; Zhou and Zhou, 2004, 2005a,b) and MUSTER by Wu and Zhang (2008).

However, since the sequence comparison algorithms [e.g., Needleman–Wunsch (Needleman and Wunsch, 1970) and Smith–Waterman (Smith and Waterman, 1981)] were proposed, gap penalty functions have much fewer developments in both information integration and performance improvement for alignment accuracy than scoring functions. The most widely used gap penalty function is the affine gap penalty (AGP), which defines the basic linear form of a gap penalty function. For a given combination of a scoring method and a linear gap penalty function,

* Corresponding author at: State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. Tel.: +86 10 6273 4376; fax: +86 10 6273 4376.
E-mail addresses: grittyy@cau.edu.cn (C. Wang), simpleyrx@163.com (R.-X. Yan), nongdaxiaofeng@126.com (X.-F. Wang), sijingna@163.com (J.-N. Si), zidingzhang@cau.edu.cn (Z. Zhang).

1476-9271/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2011.07.006

the gap penalty parameters remain fixed in aligning different residue positions. Thus, the AGP has the advantage of simplicity and easy use in dynamic programming.

There are a few strategies that have been proposed for understanding gap distributions and developing new gap penalty models. In the past two decades, several groups have attempted to find the distribution of indel lengths for a better gap penalty form (Gu and Li, 1995; Qian and Goldstein, 2001) or to empirically estimate the parameters for AGP (Benner et al., 1993; Chang and Benner, 2004; Reese and Pearson, 2002). Although there is not yet a consensus on the distribution of indel lengths to determine the optimal form of gap penalty, a few gap models have been proposed for improving the pair-wise alignment quality, including a non-local gap penalty (Taylor, 1996), generalized affine gap costs (Altschul, 1998; Zachariah et al., 2005), a "long indel" model (Miklos et al., 2004), and a logarithmic affine gap penalty (Cartwright, 2006, 2007).

Adding information derived from local structures or structure alignments to gap penalty functions has been considered to make gap placements more suitable for local sequence environments. The idea of using secondary structure information was first proposed by Lesk et al. (1986) as a variable gap penalty according to different secondary structures around the gap occurring positions. There were also several attempts to integrate statistical results from structurally aligned protein pair databases (Goonesekere and Lee, 2004; Qiu and Elber, 2006; Wrabl and Grishin, 2004). Other methods which combine environmental information into gap penalty functions include the structure-dependent gap penalty used in FUGUE (Shi et al., 2001), and the variable gap penalty scheme proposed by Madhusudhan et al. (2006).

The importance of variable gap penalties in protein sequence alignment has been demonstrated by recent progress in sequence alignment and structure prediction (Dunbrack, 2006). Profile-based variable gap penalties acquire gap information from profiles or multiple sequence alignments (MSAs) generated by PSI-BLAST search. The gap information is usually used in the form of indel frequency profiles, which is more specific for the sequences to be aligned. This kind of gap penalty schemes was previously adopted by some multiple sequence alignment programs, such as ClustalW (Thompson et al., 1994) and MAFFT (Katoh et al., 2002). At the 6th Annual International Conference on Computational Systems Bioinformatics (CSB2007), Ellrott et al. (2007) observed that using indel frequency arrays derived from PSI-BLAST MSAs could improve the alignment accuracy, especially for proteins with low sequence identity. The HHpred server (Soding et al., 2005) used the indel frequency profiles as four probabilities (insert open/extend and delete open/extend), which were converted into a position-specific gap penalty function. This idea was also employed by the SP^5 method in a different gap penalty function incorporated with the restriction of secondary structure types, although only a minor increase in alignment accuracy was achieved (Zhang et al., 2008). These above trials suggest that profile-based gap penalties could be used in a number of ways and could result in various performances of alignment accuracy.

In addition to the AGP, most of the gap models described above were published but have scarcely been used again by others, due to the difficulties in implementation and limited improvement in alignment quality. Until now, there has been no critical comparison between variable and traditional gap penalties.

In the present work, we integrated four different gap penalty models with several profile–profile scoring functions to compare them with each other. Two of the gap penalty models were linear gap penalties, the AGP and the bilinear gap penalty (BGP) (Goonesekere and Lee, 2004). The other two were profile-based variable gap penalties, the Profile-based Gap Penalty used in SP^5 (SPGP) (Zhang et al., 2008) and a new Weighted Profile-based variable Gap Penalty (WPGP) developed by us. We employed three profile–profile scoring functions, Pearson's correlation coefficient (pcc), prob_score (Mittelman et al., 2003) and prof_sim (Yona and Levitt, 2002) to measure profile vector similarities. The BLOSUM62 (b62) scoring matrix (Henikoff and Henikoff, 1992), a non-profile scoring function, was also used for comparison. Furthermore, we also compared the performances of these gap penalties with and without the secondary structure restriction (SSR).

In total, 32 different method combinations were intensively assessed in this work. For each method combination, we first optimized its parameters on a small training set, a selected version of PREFAB 4.0 (Edgar, 2004), and then tested the method with these trained parameters on several established benchmarks of different sizes [i.e., Prosup (Domingues et al., 2000), SALIGN (Marti-Renom et al., 2004) and SABmark 1.65 (Van Walle et al., 2005)]. The primary goals of this work are to determine the quantitative differences of alignment accuracy between profile-based variable gap penalties and linear gap penalties; and to establish whether there is useful gap information in profiles that could be used to improve the alignment of protein sequences.

2. Methods

2.1. Linear gap penalties

The AGP is composed of two parts: the gap opening penalty g_0 for inserting a gap and the gap extension penalty g_1 for extending each position of the gap. For a gap of length k, the penalty can be represented as

$$g(k) = g_0 + g_1 k \tag{1}$$

The BGP was suggested by Goonesekere and Lee (2004) based on the analysis of gap frequencies observed in several structurally aligned protein databases. According to the observed bilinear behavior, the BGP was defined as follows:

$$g(k) = \begin{cases} g_0 + g_1 k & (k \le 3) \\ g_0 + 3g_1 + g_2(k-3) & (k > 3) \end{cases} \tag{2}$$

where g_1 and g_2 are the extension penalties for $k \le 3$ and $k > 3$, respectively.

2.2. Profile-based variable gap penalties

The profile-based variable gap penalties that we propose here are based on the statistical indel frequency profiles from the MSAs generated by PSI-BLAST search. The indel frequency profiles were calculated as below, which is slightly different from the work of Ellrott et al. (2007) and SP^5 (Zhang et al., 2008). For each residue position i in a sequence, there are two frequencies P^i_{insert} and P^i_{delete}, which represent the probabilities of being inserted by a gap and being deleted at this position, respectively. At the corresponding position i in the MSA generated by PSI-BLAST search, P^i_{delete} is the number of sequences with a gap aligned at this position divided by the total number of sequences in the MSA. P^i_{insert} is the number of residues in the inserted gap block divided by the product of the gap length and the number of sequences in the MSA (Fig. 1). We counted every residue in the gap block but not the inserted fragments (Ellrott et al., 2007; Zhang et al., 2008) because we considered fragments of different lengths to have different probabilities of being inserted at the position. As described above, for sequences a and b to be aligned, there are four indel frequency profiles: $P^{a,i}_{insert}$, $P^{a,i}_{delete}$, $P^{b,j}_{insert}$ and $P^{b,j}_{delete}$.

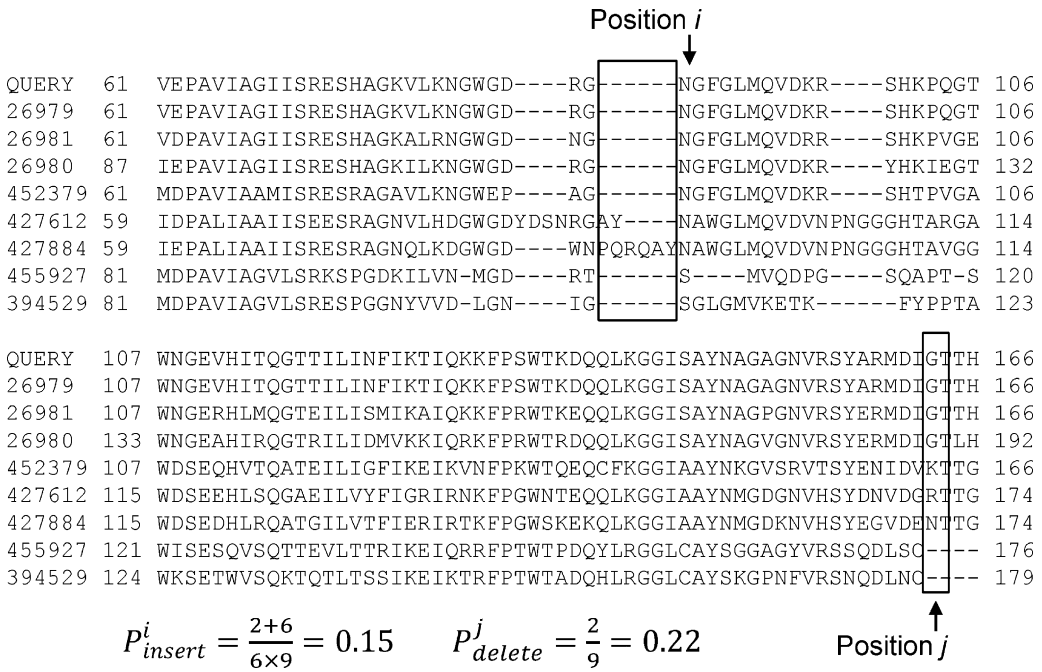

Fig. 1. Calculation of P_{insert} and P_{delete}. This is an example showing the calculation of P_{insert} of position i and P_{delete} of position j in the query sequence (see text for details).

One of the profile-based variable gap penalties implemented in this work was SPGP (Zhang et al., 2008). Our calculation of SPGP was the same as in the original paper:

$$g(k) = \begin{cases} g_0 + \sum_{j=n}^{n+k-1} g_1 \left[1 - \frac{(P^{a,i}_{insert})^\gamma + (P^{b,j}_{delete})^\gamma}{2} \right] \\ g_0 + \sum_{i=m}^{m+k-1} g_1 \left[1 - \frac{(P^{b,j}_{insert})^\gamma + (P^{a,i}_{delete})^\gamma}{2} \right] \end{cases} \quad (3)$$

where $\gamma = 0.1$ according to Zhang et al. Using the upper formula in Eq. (3), for a k-length gap inserted before position i in sequence a with k residues (from position n to $n+k-1$) aligned to this gap in sequence b, the gap penalty $g(k)$ is the opening penalty g_0 plus the sum of the extension penalties for each position j (n to $n+k-1$) of the k residues in sequence b. Similarly, the lower formula in Eq. (3) is used for the opposite situation, wherein there is a gap of length k inserted before position j in sequence b with k residues (positions m to $m+k-1$) aligned to this gap in sequence a.

In this work, we tried a new application of indel frequency profiles as another profile-based variable gap penalty called WPGP, which uses the indel frequency profiles not only in the extension penalty but also in the opening penalty. In WPGP, we further integrated a modified sequence weighting scheme (see Section 2.4) into counting the indel frequencies. The WPGP was calculated as below:

$$g(k) = \begin{cases} (g_0 + g_1) \left[1 - \frac{P^{a,i}_{insert} + P^{b,n}_{delete}}{2} \right] + \sum_{j=n+1}^{n+k-1} g_1 (1 - P^{b,j}_{delete}) \\ (g_0 + g_1) \left[1 - \frac{P^{b,j}_{insert} + P^{a,m}_{delete}}{2} \right] + \sum_{i=m+1}^{m+k-1} g_1 (1 - P^{a,i}_{delete}) \end{cases} \quad (4)$$

where the penalty for opening the first gap position (n of sequence b for the upper formula or m of sequence a for the lower formula) is the average of the corresponding non-insert and non-delete frequencies multiplied by the sum of the origin opening and extension penalties g_0 and g_1, and the penalty for every other gap extension position ($n+1$ to $n+k-1$ in sequence b or $m+1$ to $m+k-1$ in sequence a) is the origin extension penalty g_1 times the non-delete frequency of that position. The upper and lower formulae are used for different gap situations, as in SPGP.

2.3. Predicted secondary structures and SSR

Here, we used the same scheme of SSR as those in the SP-series and MUSTER algorithms (Liu et al., 2007; Wu and Zhang, 2008; Zhang et al., 2008; Zhou and Zhou, 2005a,b), in which no gaps are allowed if the secondary structure types of the current aligning position are both α-helix or both β-sheet. The secondary structure types for both sequences were predicted by PSIPRED (Jones, 1999) with all defaults against the same nrdb90 (Holm and Sander, 1998) database, which is described in the next paragraph.

2.4. PSI-BLAST profiles and sequence weights

For both profile–profile scoring functions and profile-based variable gap penalties, profiles and MSAs were constructed separately by five iterations of PSI-BLAST (version 2.2.17) (Altschul et al., 1997) searches with an E-value cutoff of 0.001. We chose to perform those searches against nrdb90 from EBI (Holm and Sander, 1998) due to its light size and fast speed. As in the work of Ohlson et al. (2004), the Position-Specific Substitution Matrices (PSSMs) directly obtained from PSI-BLAST searches were used to back-calculate the frequency profiles for the scoring functions. We directly counted the raw indel frequency profiles for SPGP, and all sequences in the MSAs were weighted equally. For WPGP, we reversed the weighting scheme in MUSTER (Wu and Zhang, 2008) to construct weighted MSAs and gap profiles from raw PSI-BLAST outputs. Sequence weights ranged from 1.0 to 0.5. Higher weights were given to sequences of higher E-values in order to amplify the effects of distantly related sequences on calculating the indel frequency profiles.

2.5. Scoring functions

We combined four different scoring functions with the aforementioned four gap penalties. They are b62, pcc, prob_score and prof_sim, which are briefly described below.

2.5.1. b62

The b62 scoring function is the only non-profile-based scoring function used in this study. The similarity score of the residue pair to be aligned is directly obtained from the BLOSUM62 matrix:

$$S_{\mathrm{b62}}(i,j) = \mathrm{BLOSUM62}(a_i, b_j) \tag{5}$$

where i and j are the positions of residues to be aligned in sequence a and b, and a_i and b_j are the corresponding residues, respectively.

2.5.2. pcc

The pcc scoring function has been previously reported and yielded a good performance on profile–profile alignment accuracy (Mittelman et al., 2003; Pietrokovski, 1996; Tomii and Akiyama, 2004; Wang and Dunbrack, 2004). Given the profile vectors a_i at position i of sequence a and b_j at position j of sequence b, the similarity score of pcc is

$$S_{\mathrm{pcc}}(i,j) = \frac{\sum_{x=1}^{20}(a_{ix}-\bar{a}_i)(b_{jx}-\bar{b}_j)}{\sqrt{\sum_{x=1}^{20}(a_{ix}-\bar{a}_i)^2\sum_{x=1}^{20}(b_{jx}-\bar{b}_j)^2}} \tag{6}$$

where x is one of the 20 amino acid types, and $\bar{a}_i$ and $\bar{b}_j$ are the average values of the vectors. The pcc scores range from -1 to 1. A large pcc score indicates the two profile vectors have similar amino acid substitution tendencies.

2.5.3. prob_score

The name "prob_score" was used in Ohlson et al.'s (2004) study of profile–profile alignment methods. The original method is PICASSO, introduced by Heger and Holm (2001, 2003) for comparing protein family profiles. Mittelman et al. (2003) modified this method into several variants. The one we chose here is PICASSO3Q, which uses only the target frequencies Q_i and Q_j from columns a_i and b_j at positions i and j of the profiles in a symmetrical equation:

$$S_{\mathrm{prob\_score}}(i,j) = \sum_{x=1}^{20} Q_{ix} \ln\frac{Q_{jx}}{p_x} + \sum_{x=1}^{20} Q_{jx} \ln\frac{Q_{ix}}{p_x} \tag{7}$$

where p_x is the background frequency of amino acid x in the database.

2.5.4. prof_sim

The prof_sim scoring method combines a divergence score and a significance score, which are calculated into a single score with the Kullback–Leibler (KL) and Jensen–Shannon (JS) divergences to measure profile similarity (Yona and Levitt, 2002). Given the target frequencies Q_i, Q_j and their average target frequency Q_0 for each type of amino acid x, the divergence score is computed as:

$$D = \frac{1}{2}\left[\sum_{x=1}^{20} Q_{ix} \log_2\frac{Q_{ix}}{Q_{0x}} + \sum_{x=1}^{20} Q_{jx} \log_2\frac{Q_{jx}}{Q_{0x}}\right] \tag{8}$$

and the significance score is calculated as:

$$S = \frac{1}{2}\left[\sum_{x=1}^{20} Q_{0x} \log_2\frac{Q_{0x}}{R_{0x}} + \sum_{x=1}^{20} p_x \log_2\frac{p_x}{R_{0x}}\right] \tag{9}$$

where R_{0x} is the average of Q_{0x} and the background frequency p_x of amino acid x.

The divergence score and the significance score are combined to produce the final similarity score:

$$S_{\mathrm{prof\_sim}}(i,j) = \frac{1}{2}(1-D)(1+S) \tag{10}$$

2.6. Dynamic programming

We used the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970), implemented by an in-house Perl script, to generate the optimal global alignment of the two sequences a and b. A shift value c was introduced to every similarity score of each scoring method, which was further trained together with the gap penalty parameters. Additionally, the terminal gaps of the alignments were not penalized by any gap penalty function.

2.7. Alignment accuracy

There are several indices available to measure the alignment accuracy by comparing the test alignments with the reference structure alignments. One index frequently used is the $Q_{developer}$ score (Q_d), which is the number of correctly aligned residue pairs (N_c) normalized by the number of all the residue pairs aligned in the reference alignment. The $Q_{modeler}$ score (Q_m) is N_c divided by the number of all pairs aligned in the test alignment. Generally, Q_d measures the sensitivity and Q_m measures the specificity of the test alignment. These two scores have been proposed or used by some groups with alternative names (Edgar and Sjolander, 2004; Marti-Renom et al., 2004; Sauder et al., 2000; Wang and Dunbrack, 2004; Yona and Levitt, 2002). We primarily used the Q_d score to measure alignment accuracy because it is only affected by N_c. We also employed other measurements such as the Cline score (Q_{Cline}) (Cline et al., 2002) and the model quality score total MaxSub (t_{MS}) (Siew et al., 2000) as complements on some of the benchmarks.

2.8. Iterative grid search for parameter optimization

To determine the optimal gap penalty parameters, an automatic iterative grid search was performed on each method. There were three parameters to be optimized for each method: the gap opening penalty g_0, the gap extension penalty g_1 and the shift value c. There was a fourth parameter, the second gap extension penalty g_2, for the BGP involved methods. Using the parameter optimization of pcc + AGP as an example, the whole optimization procedure is briefly described as follows. Regarding the AGP, three parameters (g_0, g_1 and c) needed to be optimized. We first set the initial range of these three parameters as twice the value range of the pcc scoring function. Since the value range of pcc is [−1, 1], the initial range of each parameter is [−2, 2]. Subsequently, the optimization grid search was carried out by the following steps: (i) the grid was constructed by dividing the initial range into 5 steps for each parameter; (ii) the Q_d score was calculated for each set of parameters, and the best parameter set of this round was found with the maximal Q_d score; (iii) the new range of each parameter was narrowed down centered on the best set, with 1 step range on either side, then a smaller 5-step grid was constructed; and (iv) the iteration continued, until the Q_d score of the current round was no greater than the last round, or the step was less than 0.01.

3. Results

3.1. The training results based on PREFAB 4.0

We used the same 91 protein pairs as in the SP^5 paper (Zhang et al., 2008) out of PREFAB 4.0 (Edgar, 2004) with less than 30%

Table 1
Optimized parameters of the 32 test methods by PREFAB 4.0.[a]

Method			Parameter			
SF	GP	SSR	g_0	g_1	g_2	c
b62	AGP	No	15.18	0.11	n/a	1.73
	BGP	No	11.06	2.56	0.04	0.86
	SPGP	No	14.04	1.16	n/a	0.86
	WPGP	No	13.09	1.37	n/a	−0.79
	AGP	Yes	11.29	0.80	n/a	0.19
	BGP	Yes	11.34	2.89	0.43	−0.13
	SPGP	Yes	12.36	1.16	n/a	0.72
	WPGP	Yes	11.95	2.14	n/a	−2.15
pcc	AGP	No	1.98	0.00	n/a	0.01
	BGP	No	1.35	0.43	0.19	−0.41
	SPGP	No	1.66	0.13	n/a	−0.08
	WPGP	No	1.51	0.55	n/a	−0.89
	AGP	Yes	1.14	0.04	n/a	−0.11
	BGP	Yes	1.05	0.13	0.01	0.00
	SPGP	Yes	1.78	0.12	n/a	−0.22
	WPGP	Yes	1.19	0.57	n/a	−0.87
prob_score	AGP	No	4.05	0.22	n/a	0.50
	BGP	No	3.50	0.39	0.35	0.31
	SPGP	No	4.20	0.22	n/a	0.87
	WPGP	No	3.44	0.33	n/a	0.48
	AGP	Yes	4.46	0.10	n/a	0.90
	BGP	Yes	4.08	0.44	0.33	0.38
	SPGP	Yes	3.98	0.18	n/a	0.90
	WPGP	Yes	3.52	0.59	n/a	−0.11
prof_sim	AGP	No	0.27	0.37	n/a	−1.13
	BGP	No	0.22	0.03	0.01	−0.42
	SPGP	No	0.26	0.01	n/a	−0.40
	WPGP	No	0.36	0.10	n/a	−0.58
	AGP	Yes	0.38	0.28	n/a	−0.94
	BGP	Yes	0.35	0.26	0.24	−0.87
	SPGP	Yes	0.28	0.03	n/a	−0.41
	WPGP	Yes	0.32	0.12	n/a	−0.60

[a] The same 91 pairs of PREFAB 4.0 as SP[5] were used to optimize the parameters for each method. The titles of the columns are SF – scoring function, GP – gap penalty, SSR – with or without secondary structure restriction, g_0 – gap opening penalty, g_1 –gap extension penalty, g_2 – second gap extension penalty for BGP methods, c – shift value.

Table 2
The performance of each method on the PREFAB 4.0 training dataset.[a]

Method			Performance			
SF	GP	SSR	Q_d	Q_m	Q_{Cline}	t_{MS}
b62	AGP	No	35.0	31.3	0.326	17.509
	BGP	No	36.1	32.4	0.342	17.559
	SPGP	No	35.4	31.6	0.333	17.788
	WPGP	No	**38.2**	**34.8**	**0.367**	**18.737**
	AGP	Yes	35.8	32.2	0.327	17.198
	BGP	Yes	36.3	32.6	0.343	17.694
	SPGP	Yes	36.1	32.4	0.332	17.820
	WPGP	Yes	**38.3**	**35.0**	**0.361**	**18.763**
pcc	AGP	No	53.3	47.0	0.509	24.993
	BGP	No	**54.3**	47.7	0.522	25.191
	SPGP	No	54.2	47.7	0.521	25.372
	WPGP	No	54.2	**49.3**	**0.532**	**25.538**
	AGP	Yes	54.4	48.0	0.517	25.373
	BGP	Yes	54.4	47.7	0.516	25.360
	SPGP	Yes	55.0	49.2	0.527	25.548
	WPGP	Yes	**55.8**	**49.9**	**0.541**	**25.766**
prob_score	AGP	No	53.1	47.1	0.507	24.692
	BGP	No	53.3	47.2	0.511	24.945
	SPGP	No	53.5	47.1	0.510	24.973
	WPGP	No	**53.8**	**47.8**	**0.520**	**25.139**
	AGP	Yes	53.3	46.9	0.507	24.885
	BGP	Yes	53.4	47.0	0.508	24.927
	SPGP	Yes	53.7	47.2	0.512	25.021
	WPGP	Yes	**54.1**	**48.6**	**0.522**	**25.142**
prof_sim	AGP	No	53.0	47.2	0.508	24.690
	BGP	No	53.1	47.4	0.509	24.662
	SPGP	No	53.0	47.3	0.508	24.788
	WPGP	No	**54.1**	**50.6**	**0.534**	**25.315**
	AGP	Yes	53.3	47.0	0.507	24.734
	BGP	Yes	53.7	47.2	0.510	24.752
	SPGP	Yes	53.3	47.0	0.507	24.643
	WPGP	Yes	**54.6**	**50.1**	**0.533**	**25.364**

[a] The performance on the same 91 pairs of PREFAB 4.0 as SP[5] was obtained with the optimized parameters (see Table 1). The performance scores Q_d and Q_m are shown in percentages. Total MaxSub score (t_{MS}) is the sum of MaxSub score for every test alignment of the 91 protein pairs. The best scores are shown in bold.

identity to each other to train the parameters of all 32 methods. See Table 1 for the optimized parameters of all 32 methods.

Table 2 shows the overall alignment performance of each method using the corresponding optimized parameters, quantified by Q_d, Q_m, Q_{Cline} and t_{MS}. In general, the methods using BGP, SPGP and WPGP yielded higher scores than those using AGP, and each method performed slightly better with the inclusion of SSR. The absolute increase in BGP, SPGP and WPGP compared to AGP was smaller in profile-based scoring methods than in the non-profile-based method (b62). The profile-based scoring functions achieved much higher levels of performance than did b62 no matter which gap penalty was used, indicating that modification of the gap penalty may be less powerful than the improvement from the participation of profile-based scoring functions.

3.2. The testing results based on Prosup and SALIGN benchmarks

To benchmark the 32 methods, we first tested them on two small established datasets, the Prosup (Domingues et al., 2000) and SALIGN (Marti-Renom et al., 2004) benchmarks, which were usually used as training sets in many other works. Prosup contains 127 protein pairs whose sequence alignments were generated by the structural comparison program Prosup. SALIGN consists of 200 protein pairs that share an average of 20% sequence identity. For SALIGN, we utilized the TM-align program (Zhang and Skolnick, 2005) to obtain the structural alignments as reference alignments and used TM-overlap (which has the same meaning as Q_d) and the t_{MS} score to measure the alignment accuracy and the model quality of the test alignments.

The performance on Prosup was measured by six scores: N_c, N_m, N_i, Q_d, Q'_d, and Q_m (Table 3). In terms of the Q_d score, the profile-based gap penalties (SPGP and WPGP) performed approximately 1–3% better than the linear gap penalties (AGP and BGP), except that the pcc scoring function had a <0.5% decrease. Tests on the SALIGN dataset provided similar results to those on the Prosup dataset. The performance scores, TM-overlap (Q_d), Q_m, Q_{Cline} and t_{MS}, are listed in Table 4. In general, the amount of improvement from linear to profile-based gap penalties was around 1% and the improvement brought by SSR was minor (0–1%) on both benchmark datasets. These results agree with the SP[5] method (Zhang et al., 2008), whose improvement brought by SPGP + SSR was also slight (0.5%).

3.3. The testing results based on SABmark 1.65 benchmark

Many alignment algorithms have been tested using protein pairs filtered with different criteria from the SCOP database (Murzin et al., 1995). To test the methods more intensively, we used SABmark 1.65 (Van Walle et al., 2005) as a larger benchmark set. The SABmark benchmark set was generated from the SCOP database, and it covers the entire known protein fold space with two sets: the Superfamily set and the Twilight set. The sequence identity of

Table 3
The performance of each method on the Prosup benchmark dataset.[a]

Method			Performance					
SF	GP	SSR	N_c	N_m	N_i	Q_d	Q'_d	Q_m
b62	AGP	No	5962	**1467**	10,503	36.5	53.1	25.8
	BGP	No	6138	1697	10,195	37.5	53.8	26.9
	SPGP	No	6353	1491	10,215	39.8	**56.2**	28.1
	WPGP	No	**6652**	1558	**9722**	**40.6**	55.1	**29.5**
	AGP	Yes	6344	**1512**	10,245	39.8	53.7	28.7
	BGP	Yes	6162	1646	9880	37.9	51.8	27.5
	SPGP	Yes	6542	1598	10,184	40.3	**55.6**	28.9
	WPGP	Yes	**6803**	1514	**9609**	**41.6**	54.8	**30.6**
pcc	AGP	No	9345	1035	9292	57.2	71.4	39.9
	BGP	No	**9441**	1036	9298	**57.5**	71.0	40.5
	SPGP	No	9406	1136	9307	57.1	**72.0**	40.2
	WPGP	No	9410	**975**	**8660**	57.2	70.7	**40.9**
	AGP	Yes	9420	1098	9083	57.5	**71.5**	**40.6**
	BGP	Yes	9417	1016	9389	**57.6**	71.0	40.5
	SPGP	Yes	9397	1191	**8546**	57.4	71.2	**40.6**
	WPGP	Yes	**9434**	**1009**	8828	57.1	70.9	**40.6**
prob_score	AGP	No	9386	1242	9241	57.5	71.2	40.6
	BGP	No	9382	1232	9287	57.4	71.6	40.7
	SPGP	No	9415	1105	9435	**58.0**	**72.5**	40.7
	WPGP	No	**9453**	**1070**	**9215**	57.7	71.3	**41.0**
	AGP	Yes	9312	1010	9569	57.2	71.0	39.9
	BGP	Yes	9354	1036	9504	57.3	70.3	40.1
	SPGP	Yes	9370	**1009**	9448	57.2	70.7	40.1
	WPGP	Yes	**9538**	1032	**8841**	**59.7**	**72.4**	**41.7**
prof_sim	AGP	No	9360	1291	9245	57.3	70.8	40.9
	BGP	No	9334	1409	9257	56.7	69.1	40.6
	SPGP	No	9345	1314	9228	56.9	70.1	40.6
	WPGP	No	**9454**	**1178**	**8047**	**58.6**	**71.0**	**42.7**
	AGP	Yes	9283	1089	9430	56.7	68.9	39.8
	BGP	Yes	9441	1049	9422	58.5	70.0	40.9
	SPGP	Yes	9352	1129	9385	57.1	70.3	40.2
	WPGP	Yes	**9493**	**1043**	**8255**	**58.9**	**72.3**	**42.3**

[a] The performance on the 127 protein pairs of the Prosup dataset is obtained using the optimized parameters (see Table 1) for each method. The first five columns of performance scores N_c, N_m, N_i, Q_d, Q'_d are the same as T_c, T_m, T_i, σ_0, $\sigma_{\pm 4}$ in the original Prosup paper (Domingues et al., 2000), which are the correctly aligned residue pairs (N_c), the missed pairs aligned only by the reference (N_m), the incorrect pairs aligned only by the test alignment (N_i), the average percentage of correctly aligned residue pairs (Q_d), and the average percentage of correctly aligned residue pairs within 4 positions (Q'_d), respectively. Note that there are alternative structure alignments in the files produced by Prosup, so we took the same strategy to select the best fit between the test alignment and the alternative structure alignments as the original paper, which uses the Prosup-derived structure alignment with the largest number of N_c, the lowest average shift aligned residues, and the shortest alignment length. The Q_d, Q'_d, and Q_m scores are shown in percentages with the best ones in bold.

each protein pair is lower than 25% in the Twilight set and at most 50% in the Superfamily set. The two scores f_D and f_M, which were calculated using the scripts provided by SABmark, have the same meanings as Q_d and Q_m.

The results on the SABmark 1.65 benchmark dataset showed analogous trends to those on the training set, the Prosup and SALIGN dataset. Fig. 2 shows the performance scores for each method on the Superfamily and Twilight sets of SABmark 1.65. Each score of the Superfamily set was almost twice as high as that of the Twilight set, which agrees with the corresponding sequence identity levels of the two SABmark sets. The profile-based gap penalties (SPGP and WPGP) had higher (∼2%) scores than the linear gap penalties (AGP and BGP), and most methods using SSR performed slightly better (∼1%) than the corresponding methods without SSR. The exception occurred in methods b62 + BGP and pcc + WPGP, where the performance with SSR was as good as or a bit lower (<0.3%) than the one without SSR. For prob_score and prof_sim, all of the methods using linear gap penalties (AGP and BGP) showed a <0.5% decrease in performance by integrating SSR, but the methods using profile-based gap penalties (SPGP and WPGP) with SSR performed ∼1% better than those without SSR. This result indicates that in most situations, the employment of profile-based gap penalties and SSR could yields better alignment accuracy, but the improvement is limited and inconsistent with all scoring functions.

4. Discussion

4.1. Gap distribution

To determine whether the gap penalties tested in this work fit the gap distribution of reference structure alignments, we investigated the gap distributions of all 32 methods and the reference alignments (Fig. 3). The gap distributions of references versus gap lengths are shown as dark and dashed lines, whereas those of tested methods are shown as gray and solid lines. The references' gap distributions of SALIGN and the training set differed substantially from those of SABmark. The reference lines of SALIGN and the training set indicate that there are more short gaps (length ≤3) in their reference alignments than in the test alignments, while the reference alignments of the SABmark sets contain much longer gaps. The gap distributions of the tested methods were somewhat stable in different datasets, indicating that each gap penalty function can generate gaps in alignments following the corresponding distribution. Notably, although some gap distribution analyses (Goonesekere and Lee, 2004; Gu and Li, 1995; Qian and Goldstein, 2001) have been conducted to evaluate gap penalty functions, there is no correlation between the gap distribution and the alignment accuracy. What the gap distribution could only reflect is the mathematical form of the gap penalty.

Table 4
The performance of each method on the SALIGN benchmark dataset.[a]

Method			Performance			
SF	GP	SSR	TM-overlap (Q_d)	Q_m	Q_{Cline}	t_{MS}
b62	AGP	No	32.9	31.5	29.5	41.609
	BGP	No	33.3	32.1	30.4	42.126
	SPGP	No	33.5	32.4	31.0	42.257
	WPGP	No	**34.6**	**33.7**	**32.2**	**43.532**
	AGP	Yes	32.3	31.2	28.7	41.155
	BGP	Yes	34.2	33.3	31.4	42.630
	SPGP	Yes	32.7	31.7	29.7	41.684
	WPGP	Yes	**34.3**	**33.6**	**31.6**	**44.208**
pcc	AGP	No	52.7	50.3	52.7	63.316
	BGP	No	52.9	50.6	52.8	63.425
	SPGP	No	52.9	50.5	52.8	63.681
	WPGP	No	**53.3**	**51.5**	**53.6**	**63.887**
	AGP	Yes	53.7	51.5	53.8	64.334
	BGP	Yes	53.4	51.0	53.3	63.762
	SPGP	Yes	51.8	50.3	51.8	62.634
	WPGP	Yes	**54.1**	**52.1**	**54.5**	**65.255**
prob_score	AGP	No	52.7	50.3	52.5	63.105
	BGP	No	**52.9**	50.4	52.8	63.270
	SPGP	No	51.6	49.1	51.2	61.953
	WPGP	No	**52.9**	**50.6**	**52.9**	**63.288**
	AGP	Yes	52.5	49.9	52.3	62.839
	BGP	Yes	52.5	49.9	52.4	62.879
	SPGP	Yes	52.1	49.6	51.7	62.707
	WPGP	Yes	**53.7**	**51.5**	**53.9**	**64.150**
prof_sim	AGP	No	**53.0**	50.7	**52.8**	**63.482**
	BGP	No	52.9	50.7	**52.8**	63.270
	SPGP	No	52.6	50.4	52.5	62.969
	WPGP	No	52.2	**51.6**	52.7	63.041
	AGP	Yes	52.7	50.2	52.5	63.252
	BGP	Yes	52.5	50.0	52.2	62.798
	SPGP	Yes	51.3	49.0	51.0	62.328
	WPGP	Yes	**52.9**	**51.9**	**53.5**	**63.882**

[a] The performance on the 200 protein pairs of the SALIGN dataset is obtained using the optimized parameters (see Table 1) for each method. The performance scores TM-overlap (Q_d) and Q_m are shown in percentages. Total MaxSub score (t_{MS}) is the sum of MaxSub score for every test alignment of the 200 protein pairs. The best scores are shown in bold.

4.2. The indel frequency profiles

Statistical analysis of the indel frequencies with and without sequence weights was conducted for aligned and gapped residues (which are aligned with gaps) in the reference alignments of the training set and SALIGN. As shown in Fig. 4, gapped residues tend to have larger probabilities to be inserted or deleted than aligned residues, but the difference between them is not very obvious. This implies that indel frequency profiles do carry some information to guide gap placements, but may not be good enough to represent the residue's propensity of being inserted or deleted at each position. Even though the indel frequency profiles could be amplified by the sequence weighting scheme (WPGP), according to our results, the improvement of alignment accuracy brought by WPGP is still not significant.

Ellrott et al. (2007) and Zhang et al. (2008) have used the indel frequency profiles in their gap penalty functions, but they only considered them as a part of the extension penalty. The opening penalty of an optimized gap penalty function should be much larger than the extension penalty (Table 1), and the distribution of gap lengths also shows that short gaps (e.g., length <4) are more dominant than longer gaps (Fig. 3). That is to say, the penalty of a gap is mainly determined by the opening penalty. In this work, we tried to adopt the indel frequency profiles in both the opening and extension penalties of our WPGP formula. However, the gap information captured by the indel frequency profiles from PSI-BLAST MSAs is so limited that profile-based gap penalties (SPGP and WPGP) yielded very small improvement on alignment accuracy compared to the linear gap penalties.

The probable explanations for the unexpected slight improvement of indel frequency profiles are: (i) the indel frequency profile itself is not a way good enough to capture the gap propensities from the evolutionary information in MSAs because of the small frequency difference between gapped and aligned residues and (ii) the input MSAs generated from PSI-BLAST have lots of single gaps inserted in many positions, which is too noisy to calculate the indel frequency profiles effectively. This also makes the indel frequency profiles far away from the real biological gap propensities of residues in the aligned sequences. Thus, using these indel frequency profiles, the performance of profile-based gap penalties (SPGP and WPGP) is hard to be significantly improved. Maybe that is why the development of profile–profile alignment in literature was primarily focused on the scoring function while comparatively less attention was paid on developing the gap penalty.

4.3. Secondary structure information for gaps

Some profile–profile alignment algorithms also employed the secondary structure information as one term in their scoring functions, which has proven to be helpful in improving alignment accuracy. There were also several algorithms that use secondary structure information to constrain the placement of gaps. In this work, the SSR strategy is as simple as those in SP[5] and MUSTER (Wu and Zhang, 2008; Zhang et al., 2008), which only compares the secondary structure types of the positions being aligned during the dynamic programming process. This strategy provides some help on improving the alignment accuracy according to current

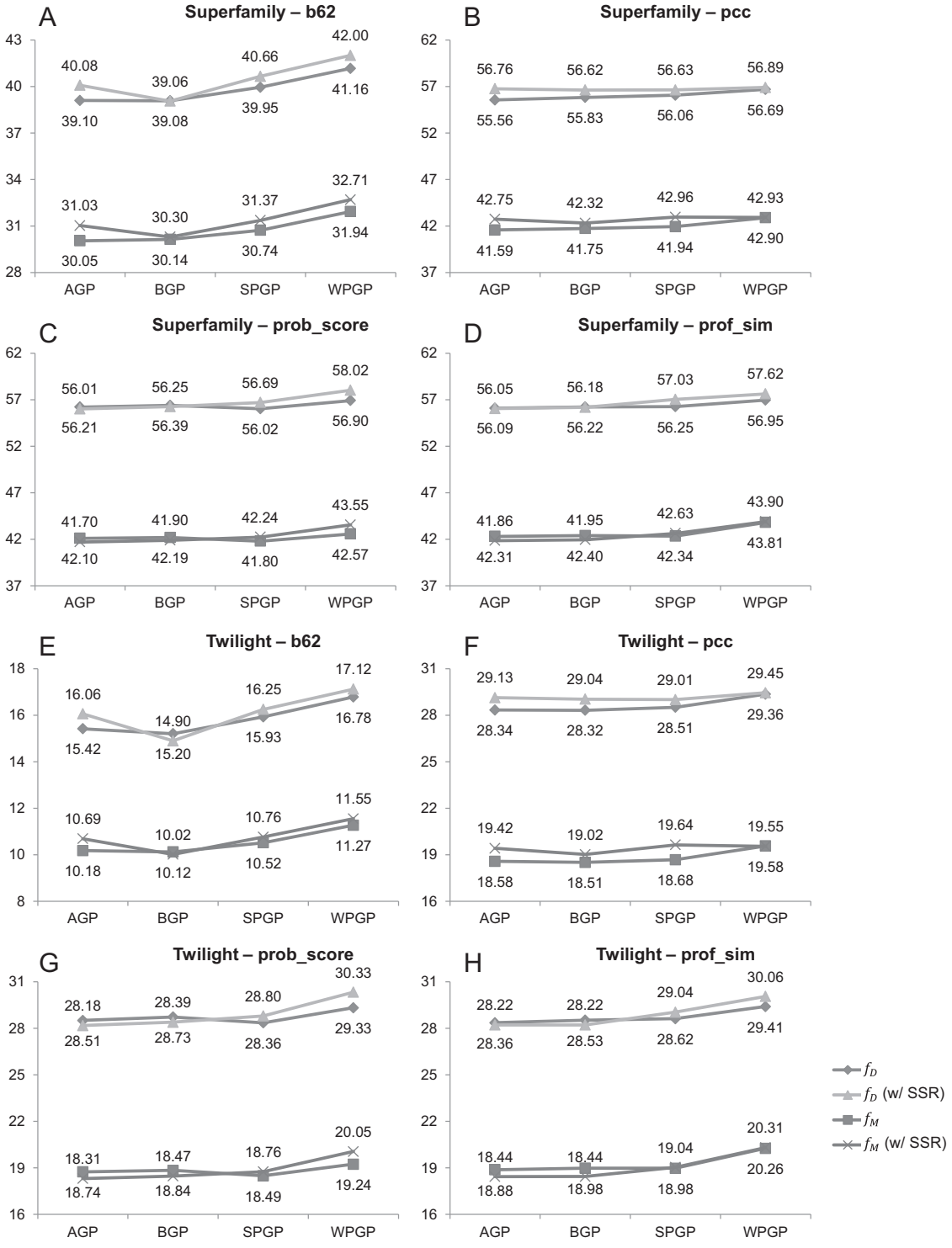

Fig. 2. The performance of each method on the Superfamily and Twilight sets of SABmark 1.65. (A–D) The performance of different scoring functions on the Superfamily set and (E–H) the performance of different scoring functions on the Twilight set. The SABmark scores f_D and f_M for all 19,092 protein pairs in 425 groups of the Superfamily set and all 10,667 protein pairs in 209 groups of the Twilight set are shown as percentages. Values with and without secondary structure restriction (SSR) are above and below the lines, respectively.

benchmark experiments. But compared to the employment of secondary structure in scoring functions of SP[5] and MUSTER, the secondary structure information appears more suitable and powerful for constructing scoring functions than being incorporated in gap penalties.

4.4. The training–testing strategy and practical use

In some alignment algorithm studies, the gap penalty parameters were trained using small training sets, whereas others only used empirical default settings like the “10 and 1” affine

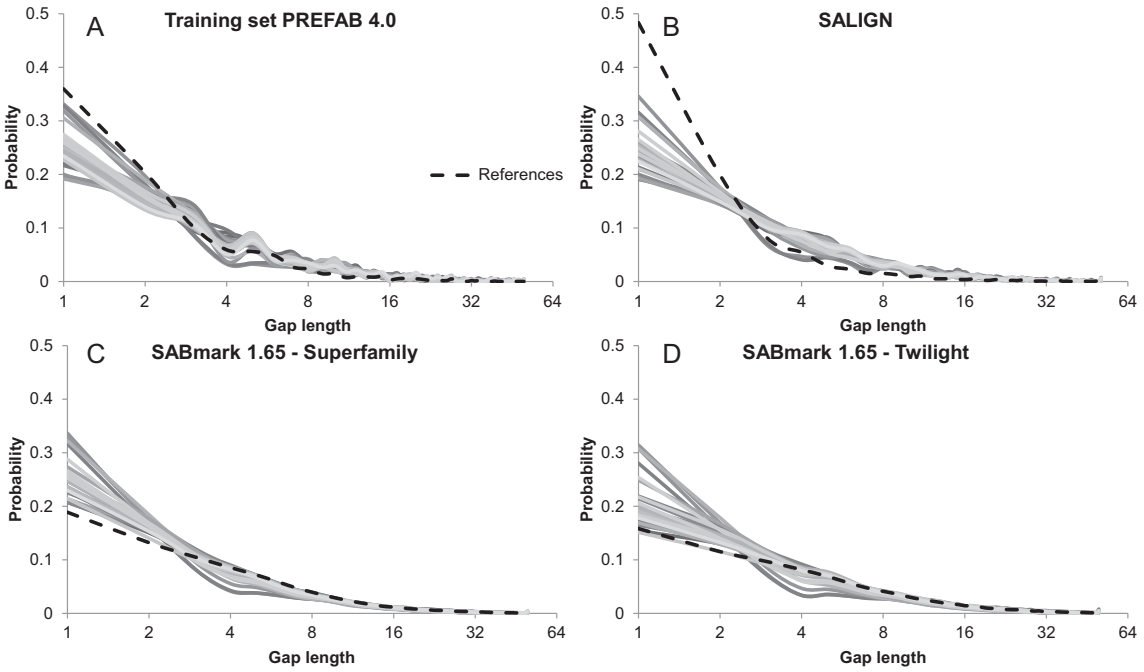

Fig. 3. Distribution of gap lengths in each dataset. (A–D) The gap distributions of all alignments generated by each method with logarithmic *X*-axes. In each panel, the gap distribution of the reference alignments is shown as a dark and dashed lines and distributions of all 32 methods are shown as gray and solid lines. The probability of each method was counted by the number of gaps of a certain length normalized by the number of gaps of all lengths in all alignments.

gap penalty for PSI-BLAST. In this work, we optimized the corresponding parameters for each gap penalty and scoring function combination by an automatic grid search optimizing approach. Our results showed that the optimization of gap penalty parameters is useful in providing better alignments, sometimes more efficient than developing a new gap penalty. Even the b62 scoring function and AGP caused a small amount of improvement in alignment accuracy after several rounds of grid search (data not shown). Therefore, it is necessary for every alignment algorithm to obtain optimized gap penalty parameters in order to maximize the performance in practical applications.

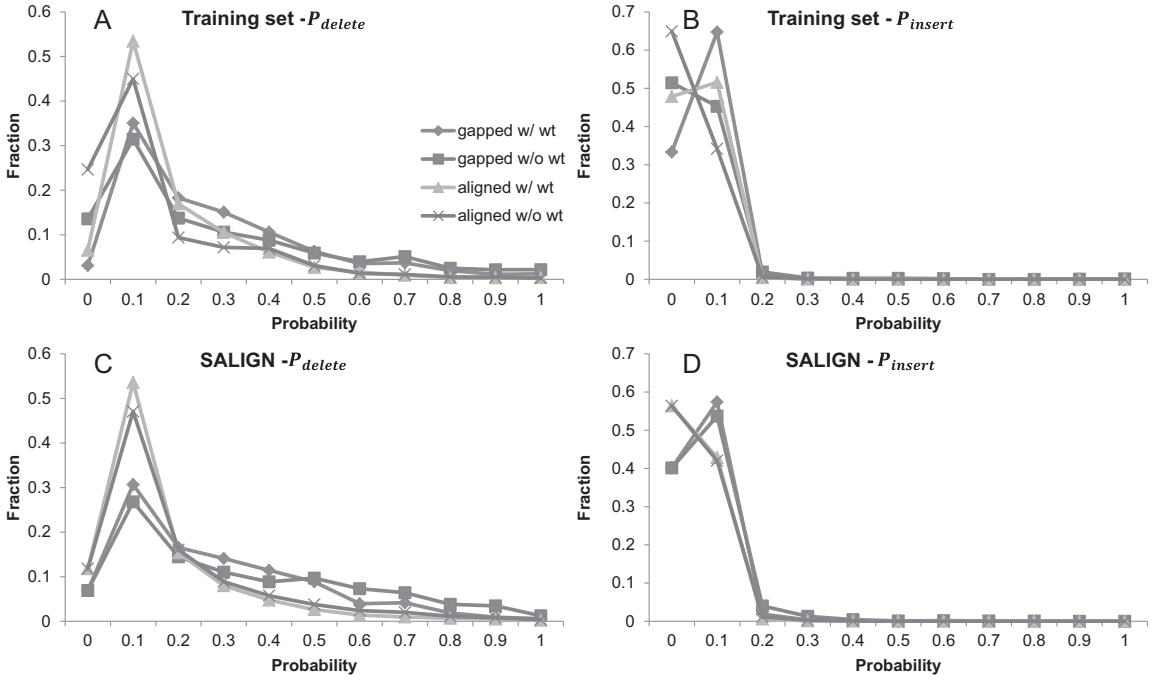

Fig. 4. Distribution of indel frequency profiles. (A–D) The distribution of indel frequency profiles (P_{insert} and P_{delete}) for the reference alignments of the training set and the SALIGN dataset. *Y*-axes are percentages of residues within the corresponding probability regions of *X*-axes. "Gapped" and "aligned" represent the residues aligned to a gap and to a residue in the other sequence, respectively.

To train and test different alignment algorithms, many training sets have been compiled with reference alignments inferred from different structure alignment programs and different similarity levels among sequences. The discrepancy between the size and quality of training sets probably makes the alignment algorithm performance results incomparable to each other. This could probably be due to the disagreement of alignments generated by different structural alignment methods. Considering that the optimization of gap penalty parameters is important but time-consuming, an ideal training set should cover the whole protein fold space within a small size. It is hoped that some "gold standard" datasets will be available in the near future to enable different alignment algorithms to be benchmarked more reliably.

5. Conclusions

In summary, we critically assessed the alignment accuracy of four different gap penalties (AGP, BGP, SPGP, and WPGP) in combining with several profile–profile scoring functions. We would like to emphasize the following findings. First, our results showed that the variable gap penalties which utilize gap information from PSI-BLAST profiles could achieve better performance than the linear gap penalties, but the overall improvement was small. Even using a proper sequence weighting scheme, the indel frequency profiles incorporated into our WPGP may not be the best way to capture biological gap information from PSI-BLAST profiles. Therefore, the maximal potential of the indel frequency profiles remains to be discovered. Second, we found that secondary structure is also beneficial information for profile–profile alignment algorithms, but when used as the SSR strategy in gap penalties, it was not as powerful as when integrated in scoring functions. Third, we found that the gap distributions cannot be effective in assessing alignment performance, although they were usually used to evaluate gap penalty functions. Finally, the optimization of gap parameters is necessary by properly selecting a well-constructed training set for both alignment algorithm assessments and practical applications. It is hoped that the current work will provide some hints for the development of new profile–profile alignment algorithms.

Acknowledgements

The authors acknowledge the support of grants from the State High Technology Development Program (2008AA02Z307) and the National Key Basic Research Project of China (2009CB918802).

References

Altschul, S.F., 1998. Generalized affine gap costs for protein sequence alignment. Proteins 32, 88–96.
Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Benner, S.A., Cohen, M.A., Gonnet, G.H., 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J. Mol. Biol. 229, 1065–1082.
Cartwright, R.A., 2006. Logarithmic gap costs decrease alignment accuracy. BMC Bioinformatics 7, 527.
Cartwright, R.A., 2007. Ngila: global pairwise alignments with logarithmic and affine gap costs. Bioinformatics 23, 1427–1428.
Chang, M.S.S., Benner, S.A., 2004. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J. Mol. Biol. 341, 617–631.
Cline, M., Hughey, R., Karplus, K., 2002. Predicting reliable regions in protein sequence alignments. Bioinformatics 18, 306–314.
Domingues, F.S., Lackner, P., Andreeva, A., Sippl, M.J., 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J. Mol. Biol. 297, 1003–1013.
Dunbrack Jr., R.L., 2006. Sequence comparison and protein structure prediction. Curr. Opin. Struct. Biol. 16, 374–384.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Edgar, R.C., Sjolander, K., 2004. A comparison of scoring functions for protein sequence profile alignment. Bioinformatics 20, 1301–1308.
Ellrott, K., Guo, J.T., Olman, V., Xu, Y., 2007. Improvement in protein sequence–structure alignment using insertion/deletion frequency arrays. Comput. Syst. Bioinformatics Conf. 6, 335–342.
Goonesekere, N.C., Lee, B., 2004. Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. Nucleic Acids Res. 32, 2838–2843.
Gu, X., Li, W.-H., 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. 40, 464–473.
Heger, A., Holm, L., 2001. Picasso: generating a covering set of protein family profiles. Bioinformatics 17, 272–279.
Heger, A., Holm, L., 2003. Exhaustive enumeration of protein domain families. J. Mol. Biol. 328, 749–767.
Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919.
Holm, L., Sander, C., 1998. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14, 423–429.
Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.
Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
Lesk, A.M., Levitt, M., Chothia, C., 1986. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. Protein Eng. 1, 77–78.
Liu, S., Zhang, C., Liang, S., Zhou, Y., 2007. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins 68, 636–645.
Madhusudhan, M.S., Marti-Renom, M.A., Sanchez, R., Sali, A., 2006. Variable gap penalty for protein sequence–structure alignment. Protein Eng. Des. Sel. 19, 129–133.
Marti-Renom, M.A., Madhusudhan, M.S., Sali, A., 2004. Alignment of protein sequences by their profiles. Protein Sci. 13, 1071–1087.
Miklos, I., Lunter, G.A., Holmes, I., 2004. A "long indel" model for evolutionary sequence alignment. Mol. Biol. Evol. 21, 529–540.
Mittelman, D., Sadreyev, R., Grishin, N., 2003. Probabilistic scoring measures for profile–profile comparison yield more accurate short seed alignments. Bioinformatics 19, 1531–1539.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.
Ohlson, T., Wallner, B., Elofsson, A., 2004. Profile–profile methods provide improved fold-recognition: a study of different profile–profile alignment methods. Proteins 57, 188–197.
Pietrokovski, S., 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res. 24, 3836–3845.
Qian, B., Goldstein, R.A., 2001. Distribution of indel lengths. Proteins 45, 102–104.
Qiu, J., Elber, R., 2006. SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. Proteins 62, 881–891.
Reese, J.T., Pearson, W.R., 2002. Empirical determination of effective gap penalties for sequence comparison. Bioinformatics 18, 1500–1507.
Sauder, J., Arthur, J., Dunbrack, R., 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40, 6–22.
Shi, J., Blundell, T.L., Mizuguchi, K., 2001. FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. 310, 243–257.
Siew, N., Elofsson, A., Rychlewski, L., Fischer, D., 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16, 776–785.
Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.
Soding, J., Biegert, A., Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244–W248.
Taylor, W.R., 1996. A non-local gap-penalty for profile alignment. Bull. Math. Biol. 58, 1–18.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.
Tomii, K., Akiyama, Y., 2004. FORTE: a profile–profile comparison tool for protein fold recognition. Bioinformatics 20, 594–595.
Van Walle, I., Lasters, I., Wyns, L., 2005. SABmark—a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 21, 1267–1268.
Wang, G., Dunbrack Jr., R.L., 2004. Scoring profile-to-profile sequence alignments. Protein Sci. 13, 1612–1626.
Wrabl, J.O., Grishin, N.V., 2004. Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. Proteins 54, 71–87.

Wu, S., Zhang, Y., 2008. MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins 72, 547–556.
Yona, G., Levitt, M., 2002. Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. J. Mol. Biol. 315, 1257–1275.
Zachariah, M.A., Crooks, G.E., Holbrook, S.R., Brenner, S.E., 2005. A generalized affine gap model significantly improves protein sequence alignment accuracy. Proteins 58, 329–338.
Zhang, W., Liu, S., Zhou, Y., 2008. SP^5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PLoS ONE 3, e2325.
Zhang, Y., Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309.
Zhou, H., Zhou, Y., 2004. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55, 1005–1013.
Zhou, H., Zhou, Y., 2005a. SPARKS 2 and SP3 servers in CASP6. Proteins 61 (Suppl. 7), 152–156.
Zhou, H., Zhou, Y., 2005b. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58, 321–328.