Position-based Sequence Weights

Steven Henikoff and Jorja G. Henikoff

Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center
Seattle, Washington 98104, U.S.A.

Sequence weighting methods have been used to reduce redundancy and emphasize diversity in multiple sequence alignment and searching applications. Each of these methods is based on a notion of distance between a sequence and an ancestral or generalized sequence. We describe a different approach, which bases weights on the diversity observed at each position in the alignment, rather than on a sequence distance measure. These position-based weights make minimal assumptions, are simple to compute, and perform well in comprehensive evaluations.

Keywords: multiple sequence alignment; sequence weighting; profiles; database searching; protein blocks

Redundancy is a common feature of sequence databanks, where a typical gene or protein family is represented by a highly non-random sample of sequences. For example, an ancient protein family might be represented by a few highly diverged microbial and invertebrate sequences plus many mammalian sequences that form a closely related subgroup. This situation can be detrimental in sequence alignment and searching applications, where it is usually desirable to represent the diversity among related sequences. Since closely related sequences are largely redundant, they provide less information in a multiple sequence alignment than their distant cousins.

Sequence weighting methods have been introduced to compensate for over-representation among multiply aligned sequences. Low weights are given to sequences that are redundant and high weights to sequences that are diverged. Sequence weights can be applied in the construction of a position-specific scoring matrix (PSSM†), such as a profile (Gribskov *et al.*, 1987), which is an ordered set of vectors, each of which represents the frequencies of residues observed for a position in a multiple alignment. By downweighting the contribution of redundant sequences to a PSSM, it should be more sensitive to distant relationships. Recent empirical results have demonstrated the value of sequence weights in increasing the sensitivity of protein sequence profiles (Thompson *et al.*, 1994; Luthy *et al.*, 1994).

While there is general agreement concerning the value of sequence weights, no consensus has been reached as to which method to use. The current methods are of two general types, tree-based and pairwise distance-based. Tree-based weights assume that sequences are related by an evolutionary tree,

and that a reasonably correct tree can be deduced from the available sequences (Felsenstein, 1985). However, this need not be the case for alignments of short and distantly related sequences, where root location can be uncertain. This uncertainty can adversely affect the ACL tree-based method (Altschul *et al.*, 1989), which upweights sequences close to the root. Uncertain root placement can cause distantly related sequences to be downweighted, and this is undesirable. For example, in the simple but non-trivial alignment of nitrogenase sequence segments shown in Table 1A, the ACL method gives zero weight to the only sequence with F in position 2, thus effectively discarding the contribution of this residue to a PSSM. To deal with the root problem, branch-proportional weights were introduced (Thompson *et al.*, 1994). This method determines the distance of each sequence from the root based on tree topology, with higher weights for sequences that share fewer nodes with other sequences. This leads to upweighting of more distantly related sequences, as desired (Table 1A). A concern with tree-based sequence weights in general is that they depend upon the particular method used for determining evolutionary distances and tree topology (e.g. see Saitou & Nei, 1987).

Pairwise distance methods (Vingron & Sibbald, 1993) do not require that sequences are related at all, and so issues such as topology and root placement are avoided. In a pairwise distance method, every sequence is assumed to lie some distance away from every other sequence, or from some generalized sequence (Vingron & Sibbald, 1993). The set of all pairwise distances for aligned sequences can be represented as a distance matrix. Two distance matrices are shown in Table 2 for the nitrogenase alignment. In the VA method (Vingron & Argos, 1989), the average number of mismatches between a sequence and the other sequences provides a measure of its distance from a hypothetical centroid. In the

574

Table 1
Comparison of different sequence weighting methods

A. *Simple but real alignment of nitrogenase segments*

Swiss-Prot ID	Residue number	Alignment	ACL	BP	VA	VOR	62%	PB
1 NIFE_CLOPA	433	GYVGS	0.250	0.282	0.269	0.259	0.250	0.267
2 NIFD_AZOVI	451	GFDGF	0.000	0.344	0.269	0.315	0.250	0.267
3 NIFD_BRAJA	461	GYDGF	0.500	0.093	0.192	0.167	0.250	0.200
4 NIFK_ANASP	475	GYQGG	0.250	0.282	0.269	0.259	0.250	0.267

B. *Contrived alignment of uniform sequences (after Sibbald & Argos, 1990)*

Alignment	ACL	BP	VA	VOR	62%	PB
AAAAA	0.333	0.167	0.188	0.177	0.167	0.167
AAAAA	0.000	0.167	0.188	0.177	0.167	0.167
CCCCC	0.333	0.167	0.188	0.177	0.167	0.167
CCCCC	0.000	0.167	0.188	0.177	0.167	0.167
TTTTT	0.333	0.333	0.250	0.292	0.333	0.333

The original Voronoi weighting scheme (Sibbald & Argos, 1990) is depicted in Tables 1 and 2 because the modified scheme (Vingron & Sibbald, 1993) is difficult to illustrate. In B, the ACL weights are practicably no different from BP, 62% or PB.

example, Sequence 1 differs from Sequences 2 to 4 at an average of 2-1/3 positions. Normalizing this distance such that all four sequence weights add up to 1 gives a VA weight of 0.269. While this method is simple and is easily calculated, it can lead to weights that are not intuitive (Sibbald & Argos, 1990; and see Table 1B).

A more elaborate pairwise distance method, Voronoi (Sibbald & Argos, 1990), supplements the observed sequences with all pseudosequences that can be derived by choosing an observed residue at each position in the alignment. Each vote of a sequence or pseudosequence is won by the observed sequence that is most similar, with equidistant sequences splitting

Table 2
Calculation of distance-based sequence weights

A. *VA sequence weights*

	GYVGS	GFDGF	GYDGF	GYQGG	Total
GYVGS	0	3	2	2	7
GFDGF	3	0	1	3	7
GYDGF	2	1	0	2	5
GYQGG	2	3	2	0	7
Total	7	7	5	7	26
Mean	7/3	7/3	5/3	7/3	26/3
Normalized	0.269	0.269	0.192	0.269	0.999

B. *Voronoi sequence weights*

Real sequences	GYVGS	GFDGF	GYDGF	GYQGG	Total
GYVGS	0 (1)	3	2	2	
GFDGF	3	0 (1)	1	3	
GYDGF	2	1	0 (1)	2	
GYQGG	2	3	2	0 (1)	
Pseudo sequences					
GYVGF	1 (1/2)	2	1 (1/2)	2	
GYVGG	1 (1/2)	3	2	1 (1/2)	
GYDGS	1 (1/2)	2	1 (1/2)	2	
GYDGG	2	2	1 (1/2)	1 (1/2)	
GYQGS	1 (1/2)	3	2	1 (1/2)	
GYQGF	2	2	1 (1/2)	1 (1/2)	
GFVGS	1 (1)	2	3	3	
GFVGF	2	1 (1)	2	3	
GFVGG	2 (1/3)	2 (1/3)	3	2 (1/3)	
GFDGS	2	1 (1)	2	3	
GFDGG	3	1 (1)	2	2	
GFQGS	2 (1/3)	2 (1/3)	3	2 (1/3)	
GFQGF	3	1 (1)	2	2	
GFQGG	3	2	3	1 (1)	
Total votes	(14/3)	(17/3)	(9/3)	(14/3)	(18)
Normalized	0.259	0.315	0.167	0.259	1.000

The columns in the distance matrices correspond to the 4 sequences in the alignment of nitrogenase segments (Table 1). In A, the rows also correspond to the 4 sequences. In B, these 4 sequences are supplemented with the 14 possible pseudosequences generated by all combinations of the residues in each position. Each entry is the number of positions where the 2 sequences differ. Each row of the distance matrix in B gets 1 vote, which is divided among the columns with the minimum value; the votes are indicated in parentheses.

Table 3
Position-based sequence weights for the alignment in Table 2A

A. Position-based residue weights

Residue	Position				
	1	2	3	4	5
G	1/(1*4)			1/(1*4)	1/(3*1)
Y		1/(2*3)			
F		1/(2*1)			1/(3*2)
V			1/(3*1)		
D			1/(3*2)		
Q			1/(3*1)		
S					1/(3*1)

B. Position-based sequence weights

Sequence	1	2	3	4	5	Total	Normalized
GYVGS	1/(1*4)	1/(2*3)	1/(3*1)	1/(1*4)	1/(3*1)	4/3	0.267
GFDGF	1/(1*4)	1/(2*1)	1/(3*2)	1/(1*4)	1/(3*2)	4/3	0.267
GYDGF	1/(1*4)	1/(2*3)	1/(3*2)	1/(1*4)	1/(3*2)	3/3	0.200
GYQGG	1/(1*4)	1/(2*3)	1/(3*1)	1/(1*4)	1/(3*1)	4/3	0.267
Total	1	1	1	1	1	5	1.001

Each residue in each position is assigned a weight equal to $1/(r*s)$ where r is the number of different residues in the position and s is the number of times the particular residue appears in the position. The position-based weights are then added for each position in each sequence.

the vote equally. In the example, the four sequences are supplemented with 14 pseudosequences (Table 2B). Sequence 2 might be considered to be the most diverged because the F in position 2 occurs where all of the other sequences have Y. This gives Sequence 2 a major advantage in competing for pseudosequences containing F at position 2. As a result, Sequence 2 captures a pleurality, whereas using the VA method, Sequence 2 is considered no more diverged than Sequences 1 and 4. Which of these methods is more correct in this example is an open question, although Vingron and Sibbald have argued that Voronoi weights are more likely to be correct in general (Vingron & Sibbald, 1993). Nevertheless, the generation of all pseudosequences becomes computationally impractical for multiple sequence alignments with many or diverse sequences, necessitating a Monte Carlo method to arrive at Voronoi weights. Even this approximate method can become impractical for larger domains (Thompson et al., 1994).

Other pairwise distance methods have provided sequence weights in special situations. For example, a clustering method was used to weight sequences within protein blocks to provide amino acid pair counts for constructing log-odds substitution matrices (Henikoff & Henikoff, 1992). Aligned sequence segment pairs that exceeded a fixed percentage of identical residues were clustered and their contributions to pair counts were averaged. For example, BLOSUM62 is the matrix derived from pair counts obtained by clustering and averaging the contributions of any segments within blocks that were more than 62% identical in pairwise comparisons. In this application, percentage identity proved to be a useful measure for generating a matrix series, in essence parameterizing the allowed degree of redundancy. This method has also been used to weight sequences within PSSMs (Henikoff et al., 1990), although the method is cruder than those described

above (e.g. see Table 1A), and the arbitrary choice of a single fixed percentage is clumsy.

A feature that tree-based and pairwise distance methods have in common is that the weight assigned to a sequence is a measure of the distance between the sequence and a root or generalized sequence. Each distance is based on the entire sequence in question. However, the sequence weights are typically applied to PSSMs in which each position vector is considered independently of all others. That is, all linkages between residues in a sequence are discarded in calculating the PSSM, and each position is considered independently when the PSSM is used. This suggests that useful sequence weights might be based on the diversity observed at each position in an alignment rather than on the diversity measured for whole sequences.

A simple method to represent the diversity at a position is to award each different residue an equal share of the weight, and then to divide that weight equally among the sequences sharing the same residue. So, if in a position of a multiple alignment, r different residues are represented, a residue represented in only one sequence contributes a score of $1/r$ to that sequence, whereas a residue represented in s sequences contributes a score of $1/rs$ to each of the s sequences. For each sequence, the contributions from each position are summed to give a sequence weight. For the nitrogenase example, the position-based weights are calculated in Table 3. Note that although Sequence 2 receives a premium for having the singular residue at position 2 as in Voronoi weights, this is balanced by penalties for having the more common residue at positions 3 and 5. In this case, the resulting weights are similar to those calculated using the VA method. However, unlike the VA method, position-based weights are always intuitively correct: in contrived examples of uniform sequences (Table 1B), such weights have been referred

to as "correct weights" (Sibbald & Argos, 1990) or "true weights" (Vingron & Sibbald, 1993) against which other methods were compared for intuitive correctness. Position-based weights are simply generalizations of correct weights from single positions to whole sequences by summation over all positions.

To assess the effectiveness of position-based weights, we carried out comprehensive evaluations. Sequence-weighted PSSMs derived from protein blocks were used to search the SWISS-PROT database (Bairoch & Boeckmann, 1992) using the PATMAT searching program (Henikoff *et al.*, 1990) and evaluated for detection of true positive sequences. This approach is analogous to that used previously to evaluate amino acid substitution matrices, in which even minor performance differences could be detected (Henikoff & Henikoff, 1993). As then, we used the PROSITE catalog of protein families (Bairoch, 1992) to provide lists of true positives for assessing performance. Blocks were constructed for all families using the fully automated PROTOMAT system (Henikoff & Henikoff, 1991). In the first series of tests, 2679 blocks representing 698 different protein groups in PROSITE 11.0 were individually tested. Every block was provided with a set of sequence weights computed according to each of the different methods described above. Blocks were then used to search SWISS-PROT 27 by conversion to sequence-weighted PSSMs. For every position of the alignment, the PSSM entry for each residue was the sequence-weighted observed frequency of that residue divided by its expected frequency tabulated from SWISS-PROT (Henikoff *et al.*, 1990).

Evaluation of searching performance was carried out by asking how many true positive sequences were detected above 99.5% of true negative sequences in the rank-ordered results list (Pearson, 1991; Henikoff & Henikoff, 1993). For each block tested, the sequence-weighted PSSMs were compared with a PSSM in which sequences were equally weighted. In most cases, all true positives were detected by all PSSMs. This is not surprising considering the demonstrated effectiveness of PSSM searches in general (Henikoff *et al.*, 1990; Henikoff & Henikoff, 1991), the fact that most of the true positives were represented in the blocks, and the fairly low detection threshold used. As a result, this test would detect performance differences only for marginal hits involving the most diverged blocks. Where there were differences between weighted and equal-weighted PSSMs representing a block, the better PSSM was considered to be that with more true positive sequences scoring above 99.5% of true negative sequences. Figure 1 (top) shows the combined results for all blocks. Every sequence weighting method provided considerably better overall performance than equal sequence weights, consistent with the empirical results of others (Thompson *et al.*, 1994; Luthy *et al.*, 1994). However, there are clear differences in performance. Three sequence weighting methods performed about equally well: position-based, Voronoi and branch-proportional weights. These provided better performance than equal
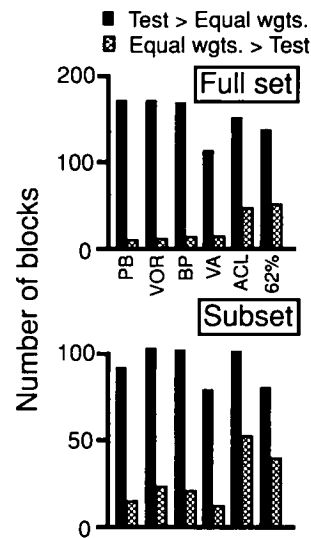
Figure 1. Evaluation of the searching performance of position-specific scoring matrices (PSSMs) using 6 different sequence weighting methods. The sequence weighting methods tested were: PB, position-based (this work); VOR, modified Voronoi (Vingron & Sibbald, 1993); BP, branch proportional (Thompson *et al.*, 1994); VA (Vingron & Argos, 1989); ACL (Altschul *et al.*, 1989); 62%, 62% clustering (Henikoff & Henikoff, 1992). BP and ACL weights for a block were calculated from the tree constructed by the Profileweight program of Thompson *et al.* (1994), which uses the neighbor-joining method (Saitou & Nei, 1987). SWISS-PROT amino acid frequencies were used to construct these PSSMs (Henikoff *et al.*, 1990). All test PSSMs were compared against a PSSM made with equal sequence weights. The filled bars represent the number of test PSSMs that scored more true positive sequences above 99.5% of the true negative sequences in the search than did the equal-weighted PSSMs. The hatched bars represent the number of equal-weighted PSSMs that scored more true positive sequences above 99.5% of the true negative sequences than did the test PSSMs. Top: Full set, using PSSMs constructed from 2679 blocks representing 698 protein groups. All true positive sequences were used to make the blocks: 2679 searches were done for each of the test PSSMs and for the equal-weighted PSSM used as a standard. Each pair of bars compares 2679 test searches with the equal sequence weighted searches. For example, 172 of the 2679 PSSMs constructed using PB sequence weights scored more true positive sequences above 99.5% of the true negative sequences than did the corresponding equal-weighted PSSMs, while 10 of the equal-weighted PSSMs did better; the PB and equal sequence weighted PSSMs performed the same in the other 2497 searches. Bottom: Subset using PSSMs constructed from 562 blocks representing 173 protein groups with at least 20 sequences. Ten true positive sequences were selected at random to make the blocks. Each pair of bars summarizes the results of 562 searches. The Profileweight program for calculating ACL and BP sequence weights was kindly provided by Julie Thompson and the Voronoi program by Peter Sibbald. Other software was written in the C programming langauge and compiled for UNIX operating systems, and is available from the authors at henikoff@howard.fhcrc.org.

sequence weights for 168 to 172 blocks and worse performance for only 10 to 14 blocks. VA weights performed less well, better than equal weights for 111

blocks and worse for 13 blocks. ACL weights also performed less well (151 to 47); 62% clustering weights performed slightly worse than ACL weights (141 to 50).

One complicating factor in this test concerns the heterogeneity of the 2679 blocks, which consist of as few as two and as many as 368 aligned sequence segments. It is possible that the performance of sequence weights depends upon the number of sequences in the block. To remove this complication, PROTOMAT was used to make blocks from a random sample of ten sequences drawn from all 173 groups with at least 20 sequences. The 562 blocks that resulted were then used to construct PSSMs for searching SWISS-PROT. Evaluation was carried out exactly as for the first test. Since half or fewer true positives were directly represented in the blocks used for searching, this second test was more challenging. In spite of these differences, the results for the second test are very similar to results for the first test (Figure 1 bottom). Position-based, Voronoi and branch-proportional weights again outperformed VA, ACL and 62% clustering weights, although VA weights show somewhat improved performance in this test. Because of the smaller number of informative groups in the second test, the results are less clear. However, a second random sample of ten sequences drawn from the 173 groups provided qualitatively similar results (data not shown).

Although our tests were carried out using ungapped protein blocks and simple PSSM scores, we expect that our conclusions will be generally applicable, for example to sequence profiles (Gribskov *et al.*, 1987) and to multiple alignments of nucleotide sequences. However, for applications that insert numerous gap characters in an alignment, gaps must be explicitly considered when using any sequence weighting scheme. Another limitation to our tests is that only a single tree construction method was used for tree-based weights, and it is possible that other methods might lead to different results.

In conclusion, the position-based approach described here provides exact and intuitively correct sequence weights directly from a multiple sequence alignment and not from an intermediary measure of distances between sequences. Since position-based sequence weights are simple to calculate and perform well, they should be appropriate for computationally demanding applications, such as iterative methods for multiple sequence alignment (Lawrence *et al.*, 1993; Krogh *et al.*, 1994; Baldi *et al.*, 1994).

References

Altschul, S. F., Carroll, R. J. & Lipman, D. J. (1989). Weights for data related by a tree. *J. Mol. Biol.* **207**, 647–653.

Bairoch, A. (1992). PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **20**, 2013–2018.

Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **20**, 2019–2022.

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994) Hidden Markov models of biolgical primary sequence information. *Proc. Nat. Acad. Sci., U.S.A.* **91**, 1059–1063.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Amer. Nat.* **125**, 1–15.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for data base searching. *Nucl. Acids Res.* **19**, 6565–6572.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 10915–10919.

Henikoff, S. & Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins: Struct. Funct. Genet.* **17**, 49–61.

Henikoff, S., Wallace, J. C. & Brown, J. P. (1990). Finding protein similarities with nucleotide sequence databases. *Methods Enzymol.* **183**, 111–132.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modelling. *J. Mol. Biol.* **235**, 1501–1531.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Luthy, R., Xenarios, I. & Bucher, P. (1994). Improving the sensitivity of the sequence profile method. *Protein Sci.* **3**, 139–146.

Pearson, W. R. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Sibbald, P. R. & Argos, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.* **216**, 813–818.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS*, **10**, 19–29.

Vingron, M. & Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *CABIOS*, **5**, 115–121.

Vingron, M. & Sibbald, P. R. (1993). Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc. Nat. Acad. Sci., U.S.A.* **90**, 8777–8781.