Weighted k-word matches: a sequence comparison tool for proteins

J. Jing[1] S. R. Wilson[2] C. J. Burden[3]

(Received 25 January 2011; revised 26 May 2011)

Abstract

The use of k-word matches was developed as a fast alignment-free comparison method for DNA sequences in cases where long range contiguity has been compromised, for example, by shuffling, duplication, deletion or inversion of extended blocks of sequence. Here we extend the algorithm to amino acid sequences. We define a new statistic, the weighted word match, which reflects the varying degrees of similarity between pairs of amino acids. We computed the mean and variance, and simulated the distribution function for various forms of this statistic for sequences of identically and independently distributed letters. We present these results and a method for choosing an optimal word size. The efficiency of the method is tested by using simulated evolutionary sequences, and the results compared with BLAST.

http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/3916 gives this article, © Austral. Mathematical Soc. 2011. Published June 4, 2011. ISSN 1446-8735. (Print two pages per sheet of paper.) Copies of this article must not be made otherwise available on the internet; instead link directly to this URL for this article.

Contents

1 Introduction

A common problem faced by biologists is to find closely related DNA or protein sequences. Sequences with a high degree of similarity are believed to be closely related in terms of evolutionary distance or to have evolved to perform functionally similar tasks. Fast algorithms are needed to search large databases to find close matches to given query sequences.

The most commonly used algorithms are based on alignments. Significance scores are attached to long alignments. These algorithms generally perform well, but fail when long range contiguity has been compromised, for example, by shuffling, duplication, deletion or inversion of extended blocks of sequence. An alternative alignment-free method is to use k-word matches, in which a significance score is attached to the number of exact matches of short words of prespecified length k [1, 2]. The algorithm for evaluating the number of k-word matches is extremely fast, with a run time linear in the lengths of the sequences being interrogated, and the method has been shown to perform at least as well as BLAST [3, 4] for simulated evolutionary DNA sequences [5].

For amino acid sequences, because of the increased alphabet size of 20 amino acids, exact matches are rare between shorter sequences. Typical optimal word sizes are two or three letters for sequence lengths below 3200 letters, and for word sizes other than optimal the accuracy decreases dramatically. We extend the idea of k-words to cumulative sums of weighted word matches. A weight is attached to k-word comparisons which acknowledges a higher rate of letter substitutions between chemically similar amino acids. Weighted word matches lead to increased optimal word sizes for shorter sequences, and the resulting scores are more stable and less sensitive to word sizes.

We compute the distributional properties of the cumulative weighted word match count and demonstrate a method for choosing an optimal word size for given sequence length. The efficiency of the method is tested by using simulated evolutionary sequences, and the results compared with BLAST. Further possible tests and future directions are also discussed.

2 Mathematical definitions and formulae

2.1 Definitions

Our starting point is the D_2 statistic, defined as the number of matches of words of prespecified length k between given sequences $\mathbf{A} = (A_1, \ldots, A_{n_A})$ and $\mathbf{B} = (B_1, \ldots, B_{n_B})$, where the letters A_i and B_j, $1 \leqslant i \leqslant n_A$ and $1 \leqslant j \leqslant n_B$, belong to some alphabet $\mathcal{A}$ of size $|\mathcal{A}| = L$. For DNA sequences, $\mathcal{A} = \{A, C, G, T\}$, and for protein sequences $\mathcal{A}$ is a set of $L = 20$ amino acids. D_2 is conveniently stated in terms of word count vectors X^A_w and X^B_w, specifying the number of times the word $w = (w_1, \ldots, w_k) \in \mathcal{A}^k$ occurs in sequences **A** and **B** respectively:

$$D_2 = \sum_{w \in \mathcal{A}^k} X_w^A X_w^B \,. \tag{1}$$

For application to protein sequences, the D_2 statistic is too stringent a measure of similarity and fails to capture the higher rate of letter substitutions between chemically similar amino acids. To account for this, D_2 is generalised to a weighted word match statistic, namely,

$$D_2^W = \sum_{w,v \in \mathcal{A}^k} X_w^A \beta_{wv} X_v^B \,. \tag{2}$$

The β-matrix represents a transition probability between words w and v and for independently evolving letters typically takes the form of a product: $\beta_{wv} = \beta_{(w_1,\ldots,w_k),(v_1,\ldots,v_k)} = \beta_{w_1 v_1} \cdots \beta_{w_k v_k}$.

The choice of transition matrix is guided by established analyses of empirical amino acid substitution rates which are encoded in the well known block substitution or BLOSUM matrices [6]. Herein we consider three ansätze for the single letter substitution matrix, or weight matrix, β_{ab}:

$$\beta_{ab}^{(1)} = q_{ab} L \,, \quad \beta_{ab}^{(2)} = q_{ab}/(p_a p_b) \,, \quad \beta_{ab}^{(3)} = q_{ab}/\sqrt{p_a p_b} \,. \tag{3}$$

Here q_{ab} is an element of the symmetric BLOSUM matrix before taking logarithms, estimated from the relative frequency of matching letter a with letter b in trusted blocks of aligned protein sequences. For each $a \in \mathcal{A}$, $p_a = \sum_{b \in \mathcal{A}} q_{ab}$ is an estimate of the relative frequency of occurrence of letter a within the blocks. In all calculations, q_{ab} was taken from the background probabilities of the BLOSUM62 matrix at the ReBLOSUM web page [7]. Although other generalisations of the D_2 statistic exist, such as D_2^* and D_2^S proposed by Reinert et al. [8], as far as we are aware, D_2^W is the first extension of the D_2 statistic to use a weight matrix with nonzero off-diagonal elements. Note that Reinert et al.'s D_2^* is a particular case of D_2^W with a diagonal β-matrix and word counts $X_w^{A,B}$ centred about their mean.

2.2 Mean and variance

Formulae for the mean and variance of the word match statistic D_2 under the assumption that the sequences $\mathbf{A}$ and $\mathbf{B}$ are composed of identically and independently distributed (iid) letters are were derived by Forêt et al. [9]. The parallel derivations for $E[D_2^W]$ and $\mathrm{Var}(D_2^W)$ are a straightforward extension, details of which are available from us. Here we simply quote results.

As in previous work [10, 9], for mathematical convenience we impose on both sequences periodic boundary conditions, that is, we define $A_i = A_{i-n_A}$, $i = n_A + 1, \dots, n_A + k - 1$, and similarly for sequence $\mathbf{B}$. The periodic boundary conditions are a minor technicality, easily implemented in practical applications. We further assume that the probability of the letter $a \in \mathcal{A}$ occurring at any given site in either sequence is f_a, where $\sum_{a\in\mathcal{A}} f_a = 1$.

We begin with the following definitions. For $a, b \in \mathcal{A}$, set

$$\eta_a = \sqrt{f_a}\,, \quad M_{ab} = \eta_a \beta_{ab} \eta_b\,, \quad \pi_t = \eta' M^{t-1} \eta\,, \quad t = 1, 2, \dots\,,$$

where the column vector $\eta = (\eta_1, \dots, \eta_L)$, and M is the $L \times L$ matrix with elements M_{ab}. We also define

$$\phi = \sum_{a,b\in\mathcal{A}} f_a f_b \beta_{ab}^2\,. \tag{4}$$

For the mean one obtains by analogy with Forêt et al. [9, Eq. (4)] the result

$$E[D_2^W] = n_A n_B \pi_2^k\,. \tag{5}$$

The derivation of $\mathrm{Var}(D_2^W)$ is nontrivial. Writing the variance as a sum of cross covariances, gives a sum of five contributions:

$$\mathrm{Var}(D_2^W) = V_1 + V_2 + V_3 + V_4 + V_5\,.$$

Analogous to work by Forêt et al. [9, Eqs. (10), (14), (17), (20) and (26)], we find

$$V_1 = n_A n_B \left(\phi^k - \pi_2^{2k}\right),$$

$$
\begin{aligned}
V_2 &= n_A n_B (n_A + n_B - 4k + 2) \\
&\quad \times \left[{\pi_3}^k + 2{\pi_2}^2 \pi_3 \frac{{\pi_3}^{k-1} - {\pi_2}^{2(k-1)}}{\pi_3 - {\pi_2}^2} - (2k-1){\pi_2}^{2k} \right], \\
V_3 &= 2 n_A n_B \left[\phi \pi_2^2 \frac{\phi^{k-1} - \pi_2^{2k-2}}{\phi - \pi_2^2} - (k-1)\pi_2^{2k} \right], \\
V_4 &= 4 n_A n_B \sum_{t=1}^{k-1} \sum_{s=0}^{t-1} \left({\pi_2}^{2s} {\pi_{2\nu+3}}^{\rho} {\pi_{2\nu+1}}^{t-s-\rho} - {\pi_2}^{2k} \right), \\
V_5 &= 2 n_A n_B \sum_{r,t=1}^{k-1} \left[\left(\prod_{i=1}^{t} \pi_{l_i} \right) \left(\prod_{j=1}^{r} \pi_{m_j} \right) - {\pi_2}^{2k} \right].
\end{aligned}
$$

In the contributions V_4 and V_5, the following definitions are used:

$$
\begin{aligned}
\nu &= \left\lfloor \frac{k-s}{t-s} \right\rfloor, \quad \rho = (k-s) \bmod (t-s), \\
l_i &= 1 + 2\eta + \left\{ \begin{array}{ll} 1 & \text{if } i \leqslant \zeta \\ 0 & \text{otherwise} \end{array} \right\} + \left\{ \begin{array}{ll} 1 & \text{if } i \leqslant \zeta - r \\ 0 & \text{otherwise} \end{array} \right\}, \\
m_j &= 1 + 2\eta + \left\{ \begin{array}{ll} 1 & \text{if } j \leqslant \zeta \\ 0 & \text{otherwise} \end{array} \right\} + \left\{ \begin{array}{ll} 1 & \text{if } j \leqslant \zeta - t \\ 0 & \text{otherwise} \end{array} \right\},
\end{aligned}
$$

where $\eta = \lfloor k/(r+t) \rfloor$, $\zeta = k \bmod (r+t)$, and $\lfloor\ \rfloor$ indicates the integer part.

3 Simulations

3.1 Distribution simulation and fitting

For each of the weight matrices in equation (3), D_2^W was simulated from a sample of $100,000$ pairs of random sequences for word lengths $k = 4, \ldots, 10$, sequence lengths n_A and n_B in the range 100 to 600, and letter frequencies $f_a = p_a$ calculated from the BLOSUM62 matrices for a 20 letter amino acid

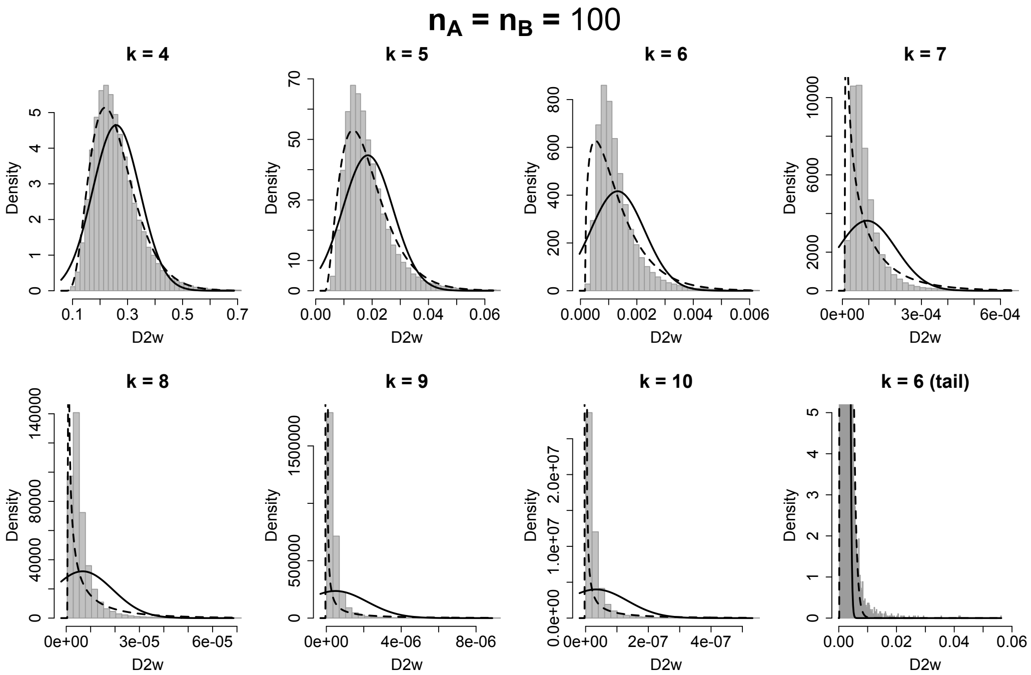

Figure 1: Simulations of D_2^W with weight matrix $\beta_{ab}^{(1)} = q_{ab}L$ from samples of $100,000$ pairs of random iid sequences for a range of word lengths k, sequence lengths $n_A = n_B = 100$, and letter frequencies $f_a = p_a$ calculated from the BLOSUM62 matrices. Superimposed are density functions for a Normal (solid curve) and Gamma (dashed curve) distribution with parameters matching the theoretical mean and variance. Also shown are details of the tail of the distribution for $k = 6$.

alphabet. The range of sequence lengths covers most single and multi-domain proteins. Examples of histograms of D_2^W are shown in Figures 1–4.

The D_2 statistic is known to be asymptotically Normal in the regime $k \ll \log n$ and asymptotically Compound Poisson in the regime $k \gg \log n$ as the

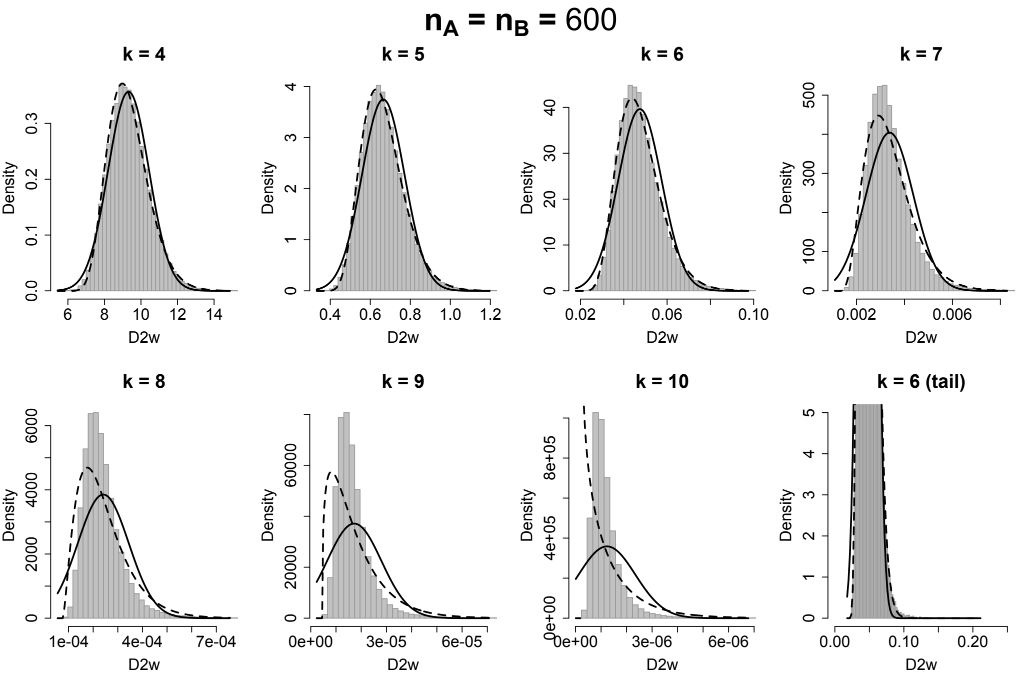

Figure 2: Simulations of D_2^W with weight matrix $\beta_{ab}^{(1)} = q_{ab}L$ from samples of $100,000$ pairs of random iid sequences for a range of word lengths k, sequence lengths $n_A = n_B = 600$, and letter frequencies $f_a = p_a$ calculated from the BLOSUM62 matrices. Superimposed are density functions for a Normal (solid curve) and Gamma (dashed curve) distribution with parameters matching the theoretical mean and variance. Also shown are details of the tail of the distribution for $k = 6$.

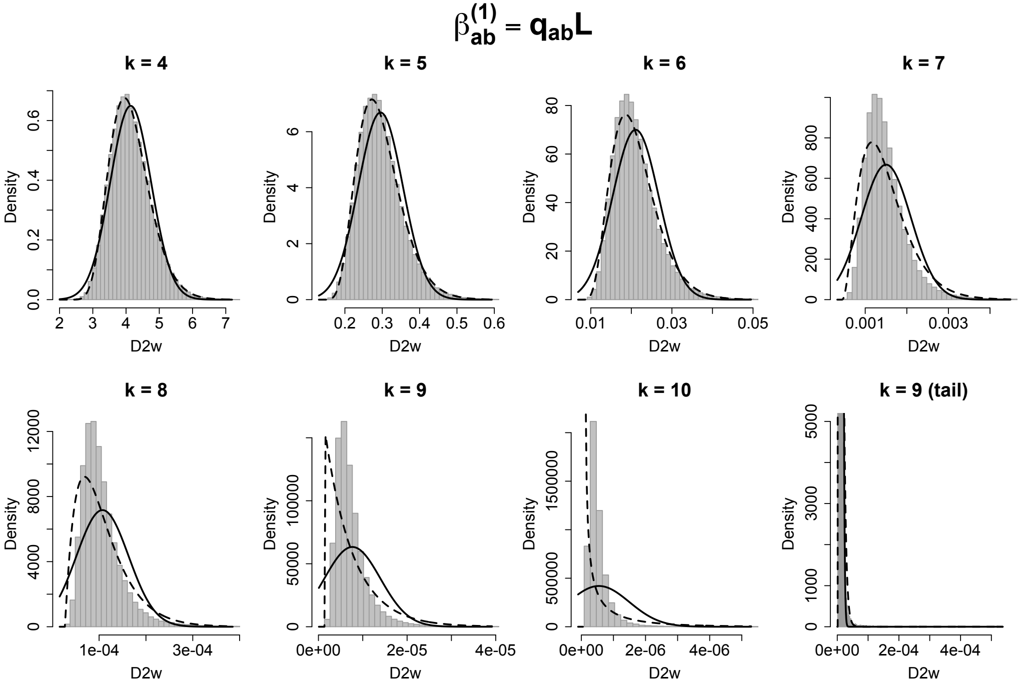

Figure 3: Simulations of D_2^W with weight matrix $\beta^{(1)}_{ab} = q_{ab}L$ from samples of $100,000$ pairs of random iid sequences for a range of word lengths k, sequence lengths $n_A = n_B = 400$ and letter frequencies $f_a = p_a$ calculated from the BLOSUM62 matrices. Superimposed density functions are as in Figure 1–2. Also shown are details of the tail of the distribution for $k = 9$.

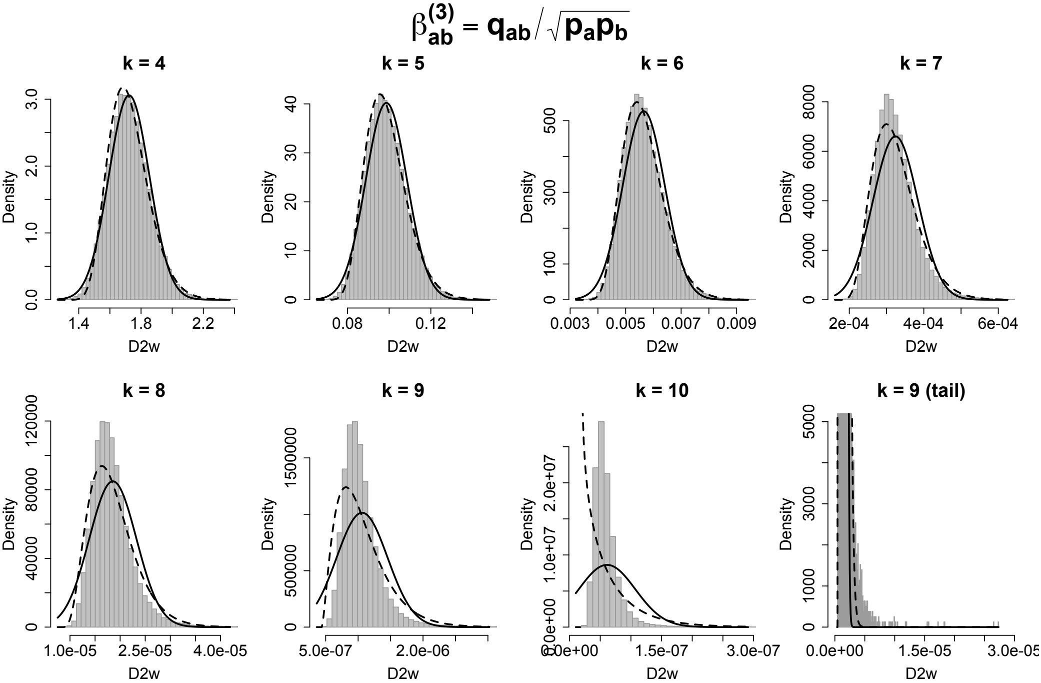

Figure 4: Simulations of D_2^W with weight matrix $\beta_{ab}^{(3)} = q_{ab}/\sqrt{p_a p_b}$, from samples of $100,000$ pairs of random iid sequences for a range of word lengths k, sequence lengths $n_A = n_B = 400$ and letter frequencies $f_a = p_a$ calculated from the BLOSUM62 matrices. Superimposed density functions are as in Figures 1–2. Also shown are details of the tail of the distribution for $k = 9$.

sequence lengths $n_A = n_B = n$ become large [1, 11]. It is also known to be well approximated by a Gamma [9] distribution in the intervening regime relevant to many biological applications. To test the appropriateness of these approximations to D_2^W we superimposed on the histograms density functions of Normal and Gamma distributions with parameter values chosen to match the above theoretical means and variances. In general, the Gamma distribution is a more accurate fit than the Normal distribution, although both distributions fail to capture the shape of the distribution at sufficiently high values of the word length k.

A transition away from a Normal distribution as the word length increases from $k <$ const. $\times \log n$ to $k >$ const. $\times \log n$, similar to that observed for D_2, is observed for D_2^W. As a rule of thumb, the distribution with weight matrix $\beta_{ab}^{(1)}$ is well represented by a Gamma distribution for $k \lesssim 2\log_{10} n$ (see Figures 1–4). Figures 3–4 show the Gamma distribution gives an improved fit for the weight matrix choice $\beta_{ab}^{(3)}$, but a poorer fit for the choice $\beta_{ab}^{(2)}$ (data not shown).

The failure of the fit at large k and smaller sequence lengths is due to the extremely long tail of the D_2^W distribution: In the final plots of Figures 1–2 an extended tail is evident for sequences of length 100, but not for sequences of length 600 when $k = 6$ and for weight matrix $\beta_{ab}^{(1)}$. A close fit can be obtained even at larger k if a Gamma distribution is fitted to the empirical distribution of $\log D_2^W$. However, this fit is of little practical use as analytic formulae of the mean and variance of $\log D_2^W$ remain intractable.

3.2 Optimal word sizes

For the D_2^W statistic to be useful as a measure of the similarity of protein sequences, it is important to establish an optimal word length corresponding to realistic biological applications. To this end, a method introduced by Wu et al. [12] using the Spearman's rank statistic to test the ability of similarity

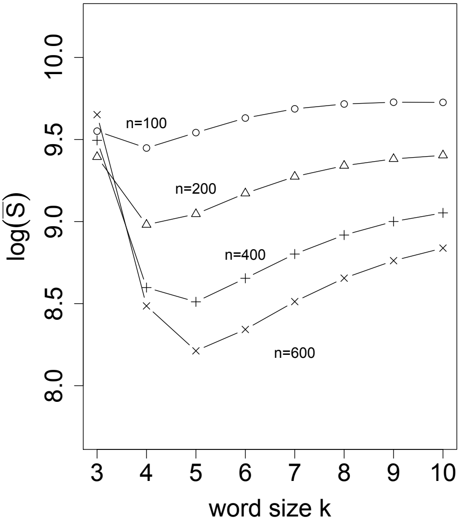

Figure 5: $\log_e$ of the Spearman's rank statistic averaged over $5,000$ families of evolved sequences for a range of sequences lengths n and word sizes k, using D_2^W with a weight matrix $\beta_{ab}^{(3)}$.

measures to estimate the relatedness of evolving sequences has previously been applied to the D_2 statistic [5].

We adapted the method to the weighted statistic D_2^W as follows: For a given sequence length $n_A = n_B = n$, we first generate a random mother sequence, then 100 generations of daughter sequences with increasing degrees of mutation using an evolution model based on amino acid substitution [13]. More specifically, each letter in the mother sequence undergoes a mutation with probability determined by a transition matrix e^{Qt}, where the matrix Q is adopted from work by Le and Gascuel [13], and the 100 daughter sequences correspond to $t = 0.01, 0.02, \ldots, 1$ respectively. The D_2^W statistic between the mother and each daughter is computed. Two rankings of the daughters are then produced, one based on generation number γ, and a ranking $r(\gamma)$ based on decreasing D_2^W. We then compute the discrepancy between these two rankings using the Spearman's rank statistic

$$S = \sum_{\gamma=1}^{100} (r(\gamma) - \gamma)^2. \tag{6}$$

Smaller S means better accuracy, and the optimal word size is defined as that for which S is minimal.

Figure 5 shows the Spearman's rank statistic averaged over 5,000 families of evolved sequences for a range of sequences lengths and word sizes using the weight matrix $\beta_{ab}^{(3)}$. The optimal word size increases with sequence length, but remains within the range for which the Gamma distribution provides an accurate approximation to the distribution of D_2^W under an iid null hypothesis.

Figure 6 compares the effectiveness of the three weight matrices defined in equation (3) and clearly indicates that $\beta_{ab}^{(3)}$ outperforms $\beta_{ab}^{(1)}$ and $\beta_{ab}^{(2)}$. Also shown for comparison is the value of the averaged Spearman's rank statistic obtained by using a ranking $r(\gamma)$ based on the BLAST P-value. There is still a gap between the performance of D_2^W and BLAST for this test. This is expected as the sequences are evolved using an amino acid substitution model which does not break sequence contiguity, and long range alignments are expected to perform best in such a situation.

4 Discussion and future work

We propose an alignment-free sequence comparison tool: the weighted k-word match D_2^W. It is designed as a measure of sequence similarity when alignments are not appropriate, and particularly when long range sequence contiguity is broken. Exact analytical formulae are given for the mean and variance of D_2^W under the iid assumption, for arbitrary alphabet, letter frequencies, sequence

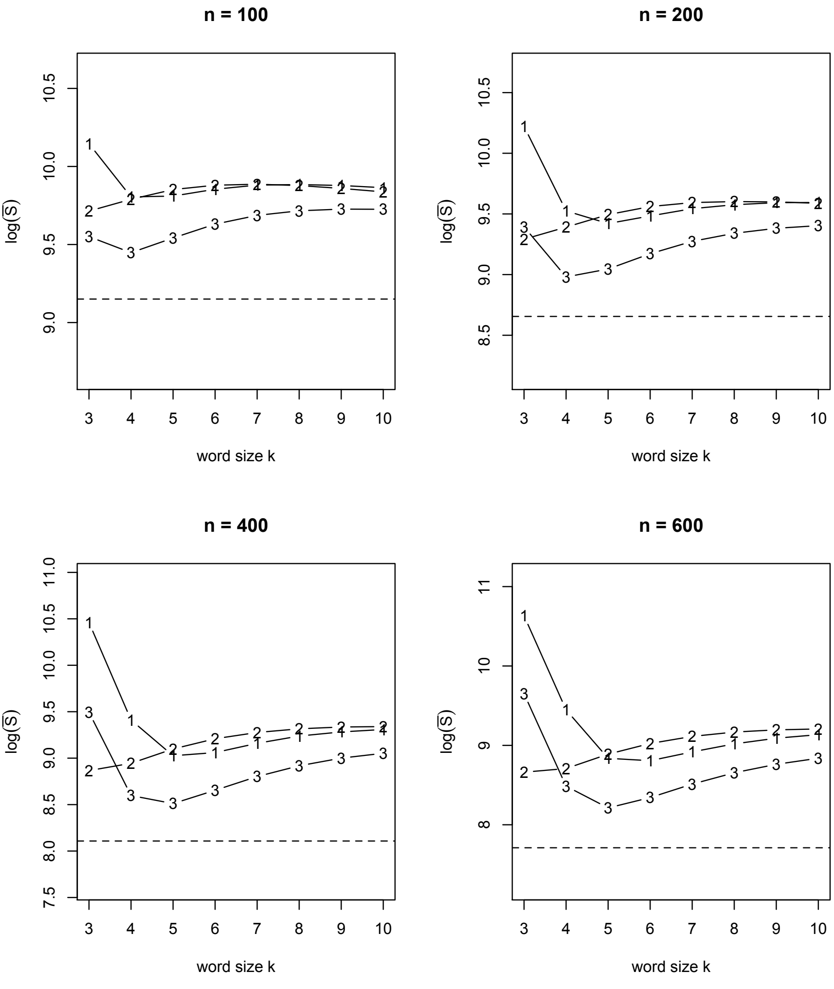

Figure 6: Comparisons of $\log_e$ of the Spearman's rank statistic for D_2^W with a weight matrices $\beta_{ab}^{(1)}$, $\beta_{ab}^{(2)}$ and $\beta_{ab}^{(3)}$. $\bar{S}$ is an average over 5,000 families of evolved sequences for a range of sequences lengths n and word sizes k. The dashed line is the value obtained from a ranking using BLAST.

lengths, weight matrix, and word size. In future work, assumptions other than iid will be considered, such as Markovian dependence or more general models using time series methods.

A number of potential weight matrices are considered, and based on amino acid evolutionary substitution rates and simulation performance, a suitable matrix chosen. For proteins in the length range 100 to 600 amino acids, which includes most proteins, the optimal word size is determined to be four to five letters using a test based on Spearman's rank statistic applied to artificially evolved sequences. For these word sizes, the distribution of the D_2^W statistics is well approximated by a Gamma distribution with parameters chosen to match the calculated mean and variance.

The true test of the D_2^W statistic is its ability to reconstruct phylogenetic trees for proteins with large numbers of shuffled or duplicated domains. Future work will apply the method to the known phylogenies of Notch receptors [14] which are part of the signal transduction pathways orchestrating cell-cell interactions and Beta-catenin [15] which plays a role in the development of embryos. These tests will benchmark further calibrating parameters to enable development of a database query tool to complement existing alignment based methods.

Acknowledgements This work was funded in part by ARC Discovery Project grant DP0987298.

References

[1] R. A. Lippert, H. Huang, and M. S. Waterman. Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. USA*, 99(22):13980–9, 2002. doi:10.1073/pnas.202468099 C173, C182

[2] J. Jing, C. J. Burden, S. Forêt, and S. R. Wilson. Statistical considerations underpinning an alignment-free sequence comparison method. *J. Korean Stat. Soc.*, 39:325–335, 2010. doi:10.1016/j.jkss.2010.02.009 C173

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–402, 1997. doi:10.1093/nar/25.17.3389 C174

[4] W. J. Ewens and G. R. Grant. *Statistical Methods in Bioinformatics: an Introduction.* Springer, 2nd edition, 2005. C174

[5] S. Forêt, M. R. Kantorovitz, and C. J. Burden. Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. *BMC Bioinformatics*, 7 Suppl 5:S21, 2006. doi:10.1186/1471-2105-7-S5-S21 C174, C183

[6] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992. doi:10.1073/pnas.89.22.10915 C175

[7] `http://bioinfo.lifl.fr/reblosum/` [31 May 2011] C175

[8] G. Reinert, D. Chew, F. Sun, and M. S. Waterman. Alignment-free sequence comparison (i): statistics and power. *J. Comput. Biol.*, 16(12):1615–1634, 2009. doi:10.1089/cmb.2009.0198 C175

[9] S. Forêt, S. R. Wilson, and C. J. Burden. Empirical distribution of k-word matches in biological sequences. *Pattern Recogn.*, 42:539–548, 2009. doi:10.1016/j.patcog.2008.06.026 C176, C182

[10] S. Forêt, S. R. Wilson, and C. J. Burden. Characterizing the D2 statistic: Word matches in biological sequences. *Stat. Appl. Genet. Mo. B.*, 8(1):Article 43, 2009. doi:10.2202/1544-6115.1447 C176

[11] M. R. Kantorovitz, H. S. Booth, C. J. Burden, and S. R. Wilson. Asymptotic behavior of k-word matches between two uniformly distributed sequences. *J. Appl. Probab.*, 44:788–805, 2006. doi:10.1239/jap/1189717545 C182

[12] T. J. Wu, Y. H. Huang, and L. A. Li. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics*, 21(22):4125–32, 2005. doi:10.1093/bioinformatics/bti658 C182

[13] S. Q. Le and O. Gascuel. An improved general amino acid replacement marix. *Mol. Biol. Evol.*, 25:1307–1320, 2008. doi:10.1093/molbev/msn067 C183

[14] E. Gazave, P. Lapébi, G. S. Richards, F. Brunet, A. V. Ereskovsky, B. M. Degnan, C. Borchiellini, M. Vervoort, and E. Renard. Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. *BMC Evol. Biol.*, 9:249, 2009. doi:10.1186/1471-2148-9-249 C186

[15] S. Q. Schneider, J. R. Finnerty, and M. Q. Martindale. Protein evolution: structure-function relationships of the oncogene Beta-catenin in the evolution of multicellular animals. *J. Exptl. Zool. (Mol. Dev. Evol.)*, 295B:25–44, 2003. doi:10.1002/jez.b.00006 C186

Author addresses

1. **J. Jing**, Mathematical Science Institute, Australian National University, Canberra, ACT 0200, Australia.
   mailto:junmei.jing@anu.edu.au

2. **S. R. Wilson**, Mathematical Science Institute, Australian National University, Canberra ACT 0200, Australia and University of New South Wales, Sydney, NSW 2052, Australia.

mailto:sue.wilson@anu.edu.au

3. **C. J. Burden**, Mathematical Science Institute, Australian National University, Canberra, ACT 0200, Australia.
mailto:conrad.burden@anu.edu.au