k-Word Matches: an Alignment-free Sequence Comparison Method

Conrad J. Burden, Sylvain Forêt, and Susan R. Wilson

Abstract— *k*-word matches, the number of words of length *k* shared between two sequences, also known as the D_2 statistic, are used in alignment-free sequence comparison statistic. The advantages of the use of this statistic over alignment-based methods for nucleotide and amino-acid sequence comparisons are firstly that it does not assume that homologous segments are contiguous, and secondly that the algorithm is computationally extremely fast, the runtime being proportional to the size of the sequence under scrutiny. We summarise our results to date on determing the distributional properties of the D_2 statistic for a range of biologically relevant parameters and outline the directions in which the research will proceed.

***Index Terms*—Genome sequences, protein sequences, sequence comparison, word matches.**

I. Introduction

ARGUABLY the most common genomic activity is sequence matching, that is, identifying segments of DNA or protein amino acid sequences which are related by identity up to genetic mutations such as insertions, deletions and point mutations. Frequently, biologists need to locate the 'closest' matches to given nucleotide or amino acid sequences from large databases of known protein sequences. The most popular, currently available sequence matching algorithms [1], such as the Basic Local Alignment and Search Tool (BLAST), rely on local alignment of the sequences being investigated, and therefore assume conservation of contiguity between homologous segments. This assumption is violated for many biologically relevant sequence comparisons. This may occur, for example, when spliced transcripts are matched to genomic sequences, when expressed sequence tags or cDNAs from different splice variants are compared, or when genomic sequences are aligned that have undergone genome shuffling.

The use of *k*-word matches [2] is an alternative sequence comparison method which overcomes these problems. It is alignment-free in the sense that it does not assume conservation of contiguity. It also has the advantage that algorithms for sequence comparisons are computationally extremely fast, being linear in the length of the query sequences.

Alignment-based database search algorithms, such as BLAST, have a sound statistical basis and provide approximate p-values and E-values (i.e. the number of high scoring matches expected under a null hypothesis). By contrast, existing implementations of database searches using *k*-word matches, such as the STACK database [3], rely on heuristics and lack any statistically rigorous measure of significance. In this paper we report on our progress in rectifying this shortcoming by extending current knowledge of the *k*-word count statistic, known as D_2, towards biologically relevant parameter regimes [4, 5, 6, 7].

II. Definitions

The D_2 statistic is defined as the number of matches of words of prespecified length k between two given sequences. Given sequences $\mathbf{A} = (A_1, \ldots, A_m)$ and $\mathbf{B} = (B_1, \ldots, B_n)$, with A_i and B_j belonging to a given alphabet,

$$D_2 = \sum_{(i,j)\in I} Y_{(i,j)},$$

where $Y_{(i,j)}$ is the *k*-word match indicator variable, equal to 1 if the word starting at position i in $\mathbf{A}$ matches the word starting at position j in $\mathbf{B}$ and 0 otherwise. The index set I is defined as $I = \{(i,j) : 1 \le i \le m - k + 1, 1 \le j \le n - k + 1\}$.

Also of use is the approximate word match statistic

$$D_2^{(t)} = \sum_{(i,j)\in I} Y_{(i,j)}^{(t)},$$

where $Y^{(t)}_{(i,j)}$, $0 \le t \le k$, is the *k*-word match indicator variable allowing up to t mismatches, i.e. $Y^{(t)}_{(i,j)} = 1$ if there are at most t mismatches between the words starting at position i in $\mathbf{A}$ and j in $\mathbf{B}$, and 0 otherwise. Clearly $D_2 = D_2^{(0)}$. Examples of exact and approximate word match counts are shown in Fig. 1. Approximate word matches have potential applications to choosing discriminative microarray probes, detection of transcription factor binding sites, microRNAs and double-stranded RNA targets, and to phylogenetics in cases where significant substitution has occurred.

III. Exact Results and Limiting Cases

To assess the significance of a particular database match, the distribution of the statistic D_2 is considered under an

C. J. Burden is jointly with the John Curtin School of Medical Research and the Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University, Canberra, A.C.T. 0200, Australia (phone: 61-2-61250730; fax: 61-2-61255549; e-mail conrad.burden@anu.edu).

S. Forêt, is with the Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University, Canberra, A.C.T. 0200, Australia (e-mail: sylvain.foret@anu.edu.au).

S. R. Wilson is with the Mathematical Sciences Institute, Australian National University, Canberra, A.C.T. 0200, Australia (e-mail: sue.wilson@anu.edu.au).

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008, Supplementary Conf. Proc., pp.235-238, 2008.

appropriate null hypothesis. The usual null hypothesis assumption is that sequences are i.i.d. strings, meaning that each letter in a sequence is independently and identically distributed. For pairs of i.i.d. strings with non-uniform letter distributions, the limiting distributions of D_2 and $D^{(t)}_2$ for large sequence lengths n can be determined for certain parameter regimes shown in Fig. 2.

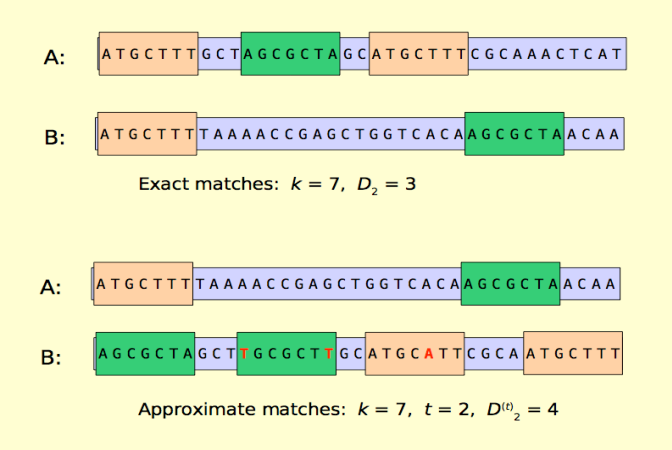

Fig. 1: Examples of exact and approximate word match counts.

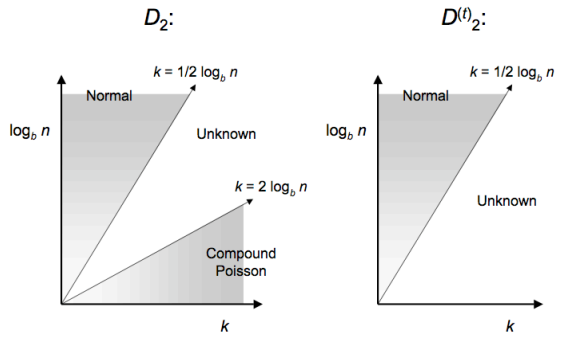

Fig. 2: Limiting distributions of exact and approximate word counts.

In the region $k < \frac{1}{2} \log_b n$ we have proved that the limiting distribution of both D_2 and $D^{(t)}_2$ as $n \to \infty$ is a Normal distribution [6]. Here the logarithm base is $b = 1/(\sum_{a \in L} f_a^2)$ with f_a the probability of occurrence of letter a in alphabet L. This result is an improvement on an the earlier result of Lippert et al. [2], who also demonstrated that the distribution of D_2 is compound Poisson in the large n limit for $k > 2 \log_b n$. No exact results are known for the limiting distribution in the intervening region between these regimes.

We have also derived exact formulae for the mean and variance of D_2 for any values of the sequence length n and word length k [5, 7]. These formulae have also been extended from i.i.d. strings to the case of Markovian strings [8], that is, sequences for which the probability of a letter occurring at a point depends on the letters immediately preceding that point.

IV. Numerical Experiments

We have carried out numerical simulations to test the accuracy with which k-word matches are able to measure the relatedness of sequences artificially evolved from an initial sequence, and to estimate the optimum word size k for given sequence length n [4].

The test, using a method similar to that introduced in [9], was carried out as follows. Beginning with an initial mother sequence of nucleotides, either randomly generated or chosen from a known genome, a family of 100 daughters were generated by randomly mutating a fraction γ% of the mother sequence, where $\gamma = 1, 2, \ldots, 100$. Only point mutations were used: substitution, insertion and deletion of a single letter. A ranking $r(\gamma)$ of the daughters was produced based on decreasing D_2 or $D^{(t)}_2$. The accuracy of the k-word match based sequence comparison was estimated by looking at discrepancy between these two rankings by means of the Spearman's rank statistic

$$A = \sum_{\gamma=1}^{100} \left(r(\gamma) - \gamma \right)^2 .$$

For each set of parameter values between 100 and 400 such families were generated and an average Spearman rank statistic determined. The optimal word size is that for which A is minimal.

Numerical results were carried out to determine the optimum word size for a range of sequence lengths n and numbers of mismatches t. Sequences with a non-uniform letter distribution were used, with nucleotide frequencies $f_A = f_T = 1/3$, $f_G = f_C = 1/6$. Similar compositional biases are observed in several sequences genomes, such as the honey bee *Apis mellifera*, the roundworm *Caenorhabditis elegans* or the zebra fish, *Danio rerio*. Results are shown in Table 1.

Table 1: Optimal word sizes for various sequence lengths n and numbers of mismatches t and a non-uniform letter distribution.

	$n = 200$	400	800	1600	3200
$t = 0$	6	7	7	7	7
1	8	10	10	10	10
2	10	12	12	12	12
3	12	14	14	14	14
4	14	16	16	16	16
5	16	18	18	18	18

Fig. 3 shows the Spearman's rank statistic for a uniform letter distribution $f_A = f_T = f_C = f_G = 1/4$. Optimum word sizes tend to be smaller for the uniform than for the non-uniform distribution. The Spearman's rank statistic increases slowly after the optimum, and the non-uniform estimates given in Table 1 might be used for most practical purposes.

Overall, the $D^{(t)}_2$ measures provide an accuracy similar to the dissimilarity measures computed in [9], with log A ranging from 9.3 to 9.6. This outperforms BLAST for this test, whose log A is close to 9.9. It is noteworthy that, at optimal word size, the $D^{(t)}_2$ statistic gives better results when the number of mismatches allowed per word increases. For this biologically relevant test we find the optimum word size generally falls within the intermediate regime in Fig. 2 for which the limiting distribution is unknown.

In the absence of analytical results in this intermediate regime we turn to numerical simulation to investigate further the distributions of D_2 and $D^{(t)}_2$. This involves generating

large samples of i.i.d. strings in the computer and comparing histograms of k-word counts with hypothesized distribution density functions.

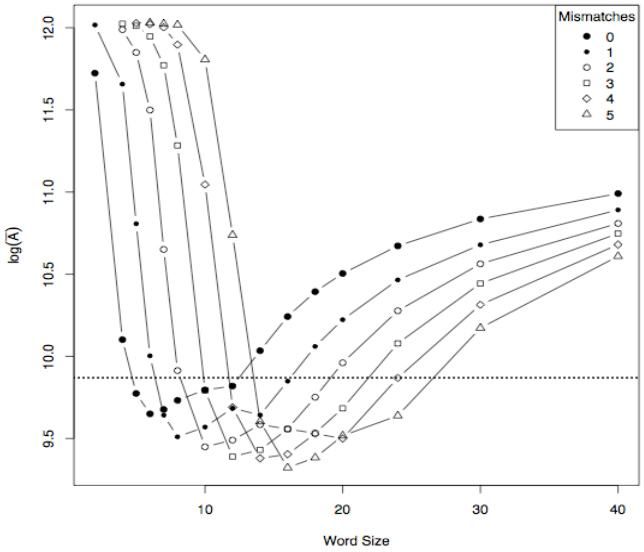

Fig. 3: The log of the average Spearman's rank statistic for a range of word sizes k and number of mismatches t, sequence size n = 600 and a uniform letter distribution. The dotted line shows the results obtained with BLAST.

We have simulated the distribution of D_2 in this way for a number of combinations of sequence size n, word size k, alphabets L and letter frequencies f_a [7]. For nucleic acid sequences, word sizes close to the optimal word size of D_2 from Table 1 were chosen. For protein amino acid sequences, the optimal word sizes and a letter composition equal to the average of the proteins encoded by the human genome were determined by the same method. For protein sequences of length up to n = 400 and an alphabet of 20 letters the optimum word size was k = 3, and for longer sequences up to n = 3200, the optimum word size was k = 4. In general, optimum word sizes decrease with increasing alphabet sizes as word matches with large alphabets are less likely. Sequences were simulated with uniform and non-uniform letter distributions, and for each combination of parameters, $N_{\text{sample}} = 10^6$ pairs of i.i.d. strings were generated.

For the purposes of estimating null hypothesis p-values, it is most important to have an accurate representation of the distribution of D_2 in the right hand tail of the distribution. Our method for evaluating a hypothesised distribution is illustrated in Fig. 4. For a range of classical significance levels, p_{hyp} = 0.001%, 0.01%, …, we calculate the corresponding quantile q_{hyp} of the hypothesised distribution, shown as the thick vertical bar in part (b) of Fig. 4. In other words, p_{hyp} is the area under the continuous curve to the right of q_{hyp}. An empirical significance level, p_{emp} is then calculated as the area of the empirical histogram (the shaded area) to the right of q_{hyp}. The discrepancy between the hypothesised and empirical p-values is then measured by the quantity

$$\delta = \log_{10}\left(p_{\text{emp}}/p_{\text{hyp}}\right).$$

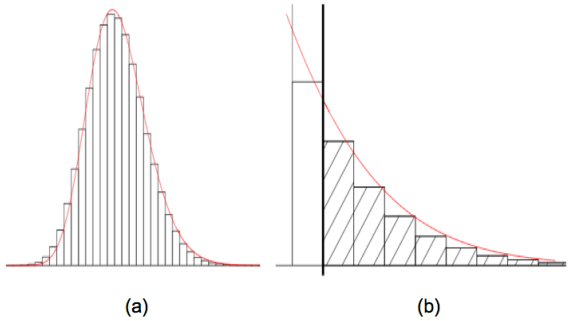

Fig. 4: The distribution of D_2 for n = 800, k = 7. The histogram shows the empirical distribution, and the continuous curve is a hypothesised Gamma distribution. (a) Global view of the distribution. (b) Detail of the right hand tail.

Values of δ close to zero indicate that the right hand tail of the hypothesised distribution is a close representation of that of the empirical distribution.

This method was used to evaluate hypothesised Normal and Gamma distributions chosen with the analytically known means and variances of the true distributions. The choice of Gamma distribution is empirical, based on the asymmetric shaped histograms of simulated samples (see Fig. 4a, for example). Tables of δ from samples of simulated DNA i.i.d. strings at empirical p-values of 1% and 0.01% are shown for the Normal distribution in Fig 5 and the Gamma distribution in Fig. 6. For sequences 1600 base pairs long or larger, the p-values from hypothesized Normal or Gamma distributions are very close to empirical p-values. For smaller sequences and p-values < 0.01%, the Normal distribution greatly overestimates the significance of D_2, while the Gamma distribution generally performs better, slightly underestimating the significance of D_2. Use of a Gamma distribution to estimate the tail of the distribution of D_2 would therefore result in fewer false positives. Identical trends were observed for amino-acid alphabets.

For data base searches, a query sequence is compared to several sequences, and the p-value of the best score of all these comparisons needs to be estimated. In this case the appropriate statistic for gauging significance of a match is the extreme value, that is, the largest of a given number of random variables. We have performed numerical estimates of the tail of the extreme value distribution of D_2, which we have compared with the extreme value distributions of the Normal and Gamma distributions using the discrepancy test described above. These two extreme value distributions belong to the Gumbel family and can be easily computed [10]. In this case we find that the Gamma distribution clearly outperforms the Normal distribution both for DNA and nucleic acid alphabets. Further details can be found in [7].

V. Conclusions and Outlook

Our investigations of the exact and approximate word match statistics indicate that p-values can be efficiently and accurately estimated for biologically relevant parameter values by assuming a Gamma distribution with mean and variance given by exact analytical results.

The above methodology has the potential to provide efficient software for alignment-free sequence matching database searches. Before this can be fully accomplished a number of directions will need to be investigated, including: (i) Different degrees of sequence similarity will be incorporated into analyses of optimum word sizes. It is well known that parameters in alignment algorithms such as BLAST must be tuned to account for the expected degree of sequence similarity. For instance, in a particular data base search, are we seeking homologous genes from closely or distantly related species? The optimum parameter values for *k*-word matches, such as word size and number of mismatches similarly must be tuned. (ii) Numerical experiments to determine the distribution of D_2 for biologically relevant parameter values will be extended from the assumption of i.i.d. strings to Markovian strings. (iii) We will evaluate whether our rigorous analytical results concerning the asymptotic distribution can be generalised to broader parameter values and to Markovian strings. (iv) Currently, the numerical algorithm for approximate mismatches is not as efficient as that for exact word matches. Improvements to the speed of this algorithm will be sought so it is also linear in the length of the query sequences, at least for the case of small numbers of mismatches.

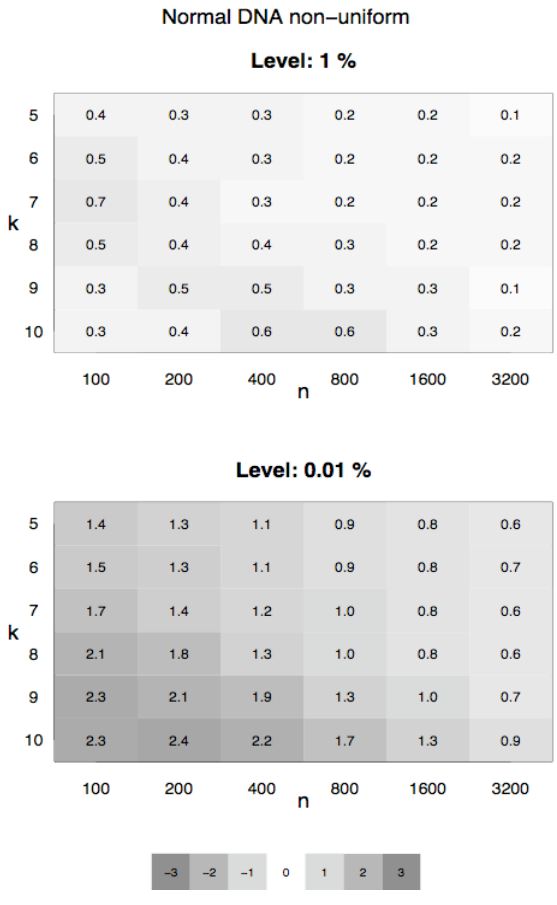

Fig. 5: The discrepancy $\delta = \log(p_{emp}/p_{hyp})$ comparing the D_2 statistic estimated from simulated DNA i.i.d. strings with a Normal distribution at empirical p-values of 1% and 0.01%.

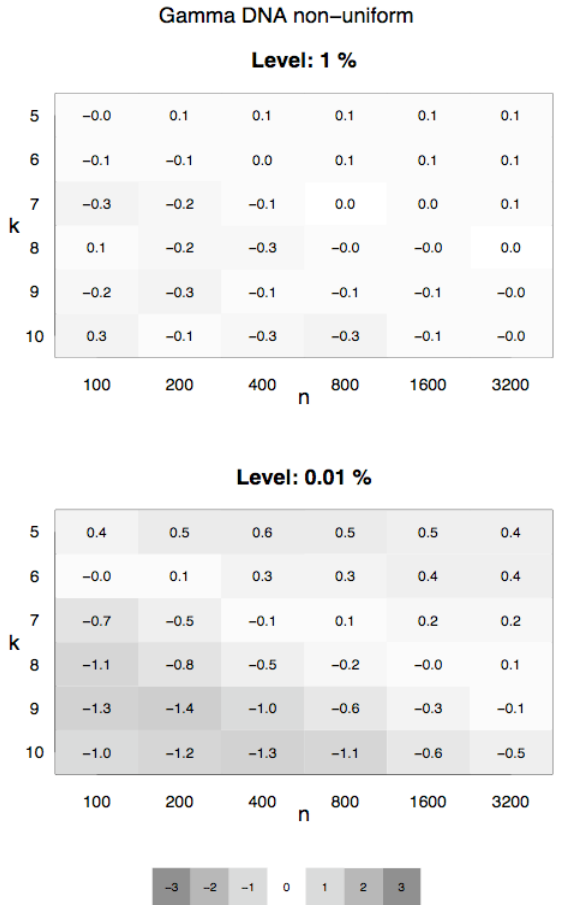

Fig. 6: The same as Fig. 5 for a Gamma distribution.

REFERENCES

[1] W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics*. New York, NY: Springer, 2nd ed. 2005.

[2] R. A. Lippert, H. Huang and M. S. Waterman, "Distributional regimes for the number of *k*-word matches between two random sequences", *Proc. Acad. Natl. Sci. USA*, vol. 99, pp. 13980–13989, 2002.

[3] A. Christoffels, et al., "STACK: Sequence tag alignment and consensus knowledge base", *Nucl. Acids Res*., vol. 29, pp. 234–238, 2001.

[4] 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*, vol. 7(Suppl 5), pp. S21, 2006.

[5] M. R. Kantorovitz, et al., "Asymptotic behaviour of *k*-word matches between two uniformly distributed sequences", *J. Appl. Prob.*, vol. 44, pp. 788–805, 2007.

[6] C. J. Burden, M. R. Kantorovitz and S. R. Wilson, "Approximate word matches between two random sequences", *Ann. Appl. Prob.*, vol. 18, pp. 1–21, 2008.

[7] S. Forêt, S. R. Wilson and C. J. Burden, "Empirical distribution of *k*-word matches in biological sequences", *Pattern Recogn*. to be published.

[8] M. R. Kantorovitz, G. E. Robinson and S, Sinha, "A statistical method for alignment-free comparison of regulatory sequences", *Bioinformatics*, vol. 23, ppi249-i255, 2007.

[9] T.J. Wu, Y. H. Huang and L. I. Li, "Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences", *Bioinformatics*, vol. 21, pp4125-4132, 2005.

[10] E. J. Gumbel, *Statistics of Extremes*. New York, NY: Columbia University Press, 1958.