BIOINFORMATICS

Vol. 19 no. 4 2003, pages 513–523
DOI: 10.1093/bioinformatics/btg005

Alignment-free sequence comparison—a review

Susana Vinga[2] and Jonas Almeida[1,*]

[1]Department of Biometry & Epidemiology, Medical University of South Carolina, 135 Cannon Street, Suite 303, PO Box 250835, Charleston, SC 29425, USA and [2]Biomathematics Group, ITQB—Universidad Nova Lisboa, PO Box 127, 2780-156 Oeiras, Portugal

Received on July 15, 2002; revised on September 27, 2002; accepted on October 6, 2002

ABSTRACT

Motivation: Genetic recombination and, in particular, genetic shuffling are at odds with sequence comparison by alignment, which assumes conservation of contiguity between homologous segments. A variety of theoretical foundations are being used to derive alignment-free methods that overcome this limitation. The formulation of alternative metrics for dissimilarity between sequences and their algorithmic implementations are reviewed.

Results: The overwhelming majority of work on alignment-free sequence has taken place in the past two decades, with most reports published in the past 5 years. Two main categories of methods have been proposed—methods based on word (oligomer) frequency, and methods that do not require resolving the sequence with fixed word length segments. The first category is based on the statistics of word frequency, on the distances defined in a Cartesian space defined by the frequency vectors, and on the information content of frequency distribution. The second category includes the use of Kolmogorov complexity and Chaos Theory. Despite their low visibility, alignment-free metrics are in fact already widely used as pre-selection filters for alignment-based querying of large applications. Recent work is furthering their usage as a scale-independent methodology that is capable of recognizing homology when loss of contiguity is beyond the possibility of alignment.

Availability: Most of the alignment-free algorithms reviewed were implemented in MATLAB code and are available at http://bioinformatics.musc.edu/resources.html.

Contact: almeidaj@musc.edu; svinga@itqb.unl.pt

INTRODUCTION

Sequence analysis is a discipline that grew enormously in recent years in response to the overwhelming burst in data generated by molecular biology initiatives. This tendency will probably continue as new challenges emerge from its quantity and increasingly integrative nature (Fuchs, 2002; Reichhardt, 1999). Although initially the algorithms were mostly borrowed from string processing computer science methodologies (Gusfield, 1997), in a second stage biological sequence analysis quickly incorporated additional concepts and algorithms from computational statistics, such as stochastic modeling of sequences using hidden Markov models and other Bayesian theory methods for hypothesis testing and parameter estimation. Both foundations carry a bias, very clear in present days, that views biological molecules as being linear sequences of discrete units similar to linguistic representations, in spite of their physical nature as a 3D structure and the dynamic nature of molecular evolution. The alignment approach overlooks well-documented long-range interactions and general fluidity resulting from recombination with shuffling of conserved segments without loss of function (Zhang *et al.*, 2002; Lynch, 2002). On the other hand, assuming conservation of contiguity allows the employment of a large set of well-developed effective computational procedures. Accordingly, the use of alignment based pair-wise sequence comparison emerges in many bioinformatic applications associated with querying a sequence database with a template, where sequence similarity is used to infer similar structure or function. Moreover, sequence divergence, leading to dissimilarity between homologous sequences, is intrinsically hard to solve as the evolutionary process takes place at different scales simultaneously (Attwood, 2000; Pearson, 2000).

The difficulty in defining a metric for sequence dissimilarity is also apparent in the analysis of natural language texts (Searls, 2001). The quantification of similarity between texts is not unique and unambiguous, depending strongly on the relative importance assigned to individual particles, letters, words, phonemes, and grammar and even to the overall context of its occurrence. The overwhelming majority of biological sequence comparison methods rely on first aligning reference homologous sequences and deriving a score for the alignment of individual units, typically the logarithm of the odds ratio. This score is

*To whom correspondence should be addressed.

then used to optimize the alignment of new sequences. Consequently, sequence dissimilarity is reduced to the comparison between candidate alignments and reference alignment of well-studied sequences, a heuristic solution for a fundamental problem which effective solution remains open. Although alignment methods are not reviewed here, comprehensive reviews abound (Durbin *et al.*, 1998; Waterman, 1995), a very brief overview of the context of its present wide use is warranted. There are two basic aspects to consider—the alignment itself and the scoring used to produce it. Optimal sequence alignment algorithms are implemented using dynamic programming, ultimately a regression technique that identifies optimal alignment by maximizing the score of the path that produces it. Several algorithms have long been identified that target specific goals such as global alignment, local alignment, with or without overlapping (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Gotoh, 1982). Although the algorithmic solutions appear satisfactory, the computational load escalates as a power function of the length of the sequences (exponent 2 for un-gapped alignment and somewhat higher for the best gapped algorithms) making its use for searching large databases unfeasible. Subsequently, a few heuristic approaches were proposed, mostly based on the recognition of alignment 'seeds', with BLAST (Altschul *et al.*, 1990, 1997) and FASTA (Pearson, 1990; Pearson and Lipman, 1988) being the most ubiquitous applications. The second critical consideration in this reference to alignment methods is the scoring of pair-wise unit alignments. A wide range of scoring systems has been proposed such as amino acid substitution scoring matrices PAM (Dayhoff *et al.*, 1978) and BLOSUM (Henikoff and Henikoff, 1992) for protein alignment. This heuristic solution reflects methodological incompleteness in the approach to sequence divergence, and also reflects assumption of conservation of contiguity between homologous segments. It is interesting to note that no scoring schemes in use will consider increasing its memory length, e.g. scoring alignment of individual oligomers rather than of individual units, equivalent to using higher order Markov model scores.

The more immediate limitations of alignment-based sequence analysis are consistently restated in all the reports reviewed below. Another difficulty, not often discussed, is that heuristic solutions make it harder to assess the statistical relevance of the resulting scores, which compromises, for example, the establishment of confidence intervals for homology. Nevertheless, the distribution of the maximum score obtained under the null hypothesis (non-correlated sequences) was deducted recently for gapped alignments (Siegmund and Yakir, 2000; Storey and Siegmund, 2001) providing a long awaited reinforcement of the theoretical foundations of scoring methods.

This report reviews published concepts and the corresponding algorithms for alignment-free comparison of biological sequences. In spite of the present surge in interest on alignment-free sequence comparison methods, there has never been, to our knowledge, any collective review of published work. However, classification, clustering or grouping techniques are not included in this overview. In cluster analysis the basic input is the cross-tabulation of dissimilarity which is then the object of agglomeration, for which there is extensive literature and widely available implementation in standard statistical packages. For a comprehensive introduction to cluster analysis and classification, see (Everitt *et al.*, 2001; Gordon, 1999). This review is confined to the measure of sequence dissimilarity itself.

BACKGROUND

The variety of disciplines involved in development of biological sequence analysis often brings together conflicting nomenclatures and conceptual frameworks. Therefore, for the convenience of the reader, some useful concepts and notation in vectors and metric spaces, information theory and word statistics are briefly recalled. References to comprehensive presentations of those fields are also included for further depth.

Words in sequences

A sequence, X, of length n, is defined as a linear succession of n symbols from a finite alphabet, A, of length r.

A segment of L symbols, with $L \leqslant n$, is designated an *L-tuple* (in some references is also defined as *L-word* or *L-plet*). The set W_L consists of all possible *L-tuples* that can be extracted from sequence X and has K elements (Equation 1).

$$W_L = \{w_{L,1}, w_{L,2}, \ldots, w_{L.K}\}$$
$$K = r^L \tag{1}$$

The identification of *L-tuples* in the sequence X can then be the object of counting occurrences with overlapping (Equation 2). Computationally, the counting is usually performed by taking a sliding window L-wide that is run through the sequence, from position 1 to $n - L + 1$.

$$c_L^X = (c_{L,1}^X, \ldots, c_{L,K}^X) \tag{2}$$

Similarly, one can then calculate word frequencies, f_L^X, to estimate the probability, $p_{L,i}^X$, of finding a specific word $w_{L,i}^X$, collectively defining a vector of word or *L-tuple* probabilities (Equation 3).

$$p_L^X = (p_{L,1}^X, p_{L,2}^X, \ldots, p_{L,K}^X) \tag{3}$$

The vector of frequencies f_L^X is obtained as the relative abundance of each word (Equation 4).

$$f_L^X = \frac{c_L^X}{\sum_{j=1}^{K} c_{L,j}^X} \Leftrightarrow f_{L,i}^X = \frac{c_{L,i}^X}{n - L + 1} \tag{4}$$

For example for DNA sequences, $A = \{A, T, C, G\}$, $r = 4$, a three letter word, $L = 3$, could be $w_3 = ATC$. For the sequence $X = ATATAC$, where $n = 6$, the vector p_3^X is estimated by the relative frequencies of all trinucleotides. The frequencies, determined by sliding a three letter window $n - L + 1 = 4$ times would be:

$$W_3 = \{ATA, TAT, TAC, AAA, \ldots\}$$
$$c_3^X = (2, 1, 1, 0, \ldots)$$
$$f_3^X = (0.5, 0.25, 0.25, 0, \ldots)$$

The vectors c_3^X and f_3^X have length $K = 4^3 = 64$ and the zero coordinates correspond to missing words in X, in this case absent trinucleotides.

Distance between sequences

A distance function $d(X, Y)$ is a function that assigns a real number to every pair X and Y belonging to a given set, in this application will be the set of all possible sequences. In order for $d(X, Y)$ to be a metric distance (Strang, 1988) the three properties in Equation (5) have to be observed.

$$\begin{aligned} &\text{Positivity} : d(X, Y) \geq 0 \text{ and } d(X, Y) = 0 \Leftrightarrow X = Y \\ &\text{Symmetry} : d(X, Y) = d(Y, X) \\ &\text{Triangle inequality} : d(X, Y) + d(Y, Z) \geq d(X, Z) \end{aligned} \tag{5}$$

Most of the distance functions reviewed below are computed in the spaces defined by the vectors of word counts and word frequencies. For a comprehensive introductory study of linear algebra and vector spaces see Strang (1988) and for an introduction to matrix analysis Schott (1997) is recommended.

Word statistics

The statistical and probabilistic properties of words in sequences were recently systematized and reviewed (Reinert *et al.*, 2000), with emphasis on the deductions of exact distributions and the evaluation of its asymptotic approximations. The problems addressed in that report included finding formulae for counts expectation, variances and also covariances between frequencies of two words, namely the distribution of p_L^X and the determination of its moments. These issues are fundamental to assess the statistical significance of dissimilarity results based on frequencies of words. The period or overlap capability deserves special mention here, as it will be of importance for the reviewing, below, of metrics based on the Mahalanobis distance. It indicates to what extent the prefix and the suffix of a word are equal, i.e. if the word beginning is the same as the ending (Gentleman and Mullin, 1989). This property is fundamental to the correct deduction of the covariances of p_L^X, as words that share motifs are more likely to co-occur. The modeling of the resulting word statistics is often approached within the framework of the theory of stochastic processes, namely Markov chains and renewal theory (Gentleman and Mullin, 1989; Régnier, 1998; Reinert *et al.*, 2000; Waterman, 1995) and will not be reviewed here.

Information theory

Information theory was originated in the classical paper of Claude Shannon in 1948 (Shannon, 1948) to quantify the capability of transmitting data over a channel. Some years later, Solomon Kullback formalized it as a branch of statistical theory (Kullback, 1968) and gave rigorous mathematical proofs of theorems previously introduced. The main concept behind information theory is the notion of entropy or uncertainty. One defines the entropy of a random variable based on the probabilities of all the outcomes. The definition will be subsequently applied to sequences, where the random variable represents an *L-tuple*. The entropy H of *L-tuples*, W_L, is calculated from the probability of the individual words in sequence X (Equation 6).

$$H(W_L^X) = -\sum_{i=1}^{K} p_{L,i}^X \log_2(p_{L,i}^X) \tag{6}$$

This general definition is valid for any word length resolution, L, including the more common determination of uncertainty associated to the distribution of individual symbols, e.g. by using $L = 1$. It was subsequently shown that this is the only function that satisfies some logically required axioms for the quantification of uncertainty (Ash, 1990), such as addictively of entropies for joint probability spaces, the fact that $H(W)$ is maximal when all the K possible words are equiprobable, $H(W) = \log_2(K)$, and it is minimal when $p_{L,i}^X = 1$ for some i-word—knowing the outcome should make uncertainty equal to zero. $H(W)$ is also an increasing function of K equiprobable spaces, i.e. it will be higher if the number of possible words increase. Comprehensive presentations of this matter and respective applications abound, such as Cover and Thomas (1991). For the studies reviewed below it is useful to detail the Kullback–Leibler (KL) discrepancy, measuring relative entropy between two discrete probability distributions p and q, detailed in Equation (7).

$$KL(p, q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \tag{7}$$

However, it is noteworthy that the KL discrepancy is not a metric distance because it only satisfies positivity but not symmetry nor triangular inequality (Equation 5).

ALIGNMENT-FREE SEQUENCE COMPARISON

The proposition of alignment-free methods to compare biological sequences is a very recent endeavor, with the earliest systematic publications being less than 2 decades old (Blaisdell, 1986). Although the pace of work in this area is increasing sharply, the total number of published reports proposing or using alignment-free metrics is relatively small, still under the one hundred mark. Moreover, the past decade contains the overwhelming majority of reports and judging by those published in the past year, the trend is being maintained. Two main categories of proposed methods can be recognized in the literature reviewed—methods based in word frequency, and those that do not require resolving the sequence with fixed word-length segments. The first group includes procedures based on metrics defined in coordinate space of word-count vectors, such as the Euclidean distance and relative entropy of frequency distributions. On the contrary, the second category corresponds to techniques that are independent from the resolution of the sequence, i.e. they do not involve counting segments of fixed length. They include the use of Kolmogorov complexity theory and scale-independent representation of sequences by iterative maps. These two categories of methods have distinct theoretical lineages and an unequal amount and variety of techniques explored in the published reports, far fewer for the latter.

Methods based on word frequencies

All methods described in this section start with the mapping of sequences to vectors defined by the counts of each *L-tuple*. This straightforward approach was the first attempt to transform a sequence into an object for which Linear Algebra and Statistical Theory had useful analytical tools already available. The vectors obtained represent the original sequence with a fixed resolution L, that of the word length considered. The basic rationale for sequence comparison is that similar sequences will share word composition to some extent, which is then quantified by a variety of techniques. This is, in a way, an extension of the widespread use of difference in GC content as a measure of sequence dissimilarity. It is noteworthy that the methods described here, although alignment-free, are still length dependent in the sense that the comparisons are made for fixed word length. This could even be viewed as a weak departure from the idea of alignment since sharing *L-tuples* is equivalent to recognizing an alignment between identical segments. However, a variety of methods have been proposed to derive combined distance metrics that contain information about all resolutions, in order to achieve complete independence from the contiguity of conserved segments.

EUCLIDEAN DISTANCE

The first published report systematizing the use of *L-tuple* counts for sequence comparison dates from 1986 (Blaisdell, 1986). In this work, the author presents a new measure of dissimilarity between sequences modeled as Markov chains. The difference between two sequences was quantified by the square Euclidean distance between their transition matrices. In spite of its conceptual simplicity, this method was shown to be an effective alternative to alignment methods. The fact that a transition matrix of a Markov chain can be identified with the frequency of all *L-tuples* lead the author to propose other quantifications of sequence similarity, such as the use of Chi-square tests to assess the statistical significance of a specific comparison (Blaisdell, 1986). It was further shown in this pioneering report that the approach enabled the measure of dissimilarity between sequences that are too different to be amenable to alignment, even if they still have recognizable similarity. The fact that, when alignment is possible the two methods agree, provided further support for the adoption of the more generally applicable alignment-free alternative. For each resolution or word-length L, the squared Euclidean distance between sequences X and Y is determined by Equation (8), where $c_L^X = (c_{L,1}^X, \ldots, c_{L,K}^X)$ and $c_L^Y = (c_{L,1}^Y, \ldots, c_{L,K}^Y)$ are vectors representing word counts for those sequences and K is the number of different *L-tuples* possible for that L-length.

$$d_L^E(X,Y) = (c_L^X - c_L^Y)^{\mathrm{T}} \cdot (c_L^X - c_L^Y) = \sum_{i=1}^{K} (c_{L,i}^X - c_{L,i}^Y)^2 \tag{8}$$

Nevertheless, alignment was still observed in the same report to be more accurate for comparison of sequences with very close similarity. A few years later the same author formalized the new alignment-free metric and validated its performance by successfully comparing large genomic sequences from organisms with well documented phylogenetic relationships (Blaisdell, 1989b). The dissimilarity values obtained by pair-wise sequence comparison was subsequently used to correctly recognize phylogenetic relationships with PHYLIP package (Felsenstein, 1993), corroborating results obtained with 'conventional methods that assume prior correct homologous total alignment of the sequences'. A similar conclusion was reached in a subsequent study (Blaisdell, 1989a) where the dissimilarity values obtained with alignment-free Euclidean distance were observed to be directly proportional to conventional mismatch counts requiring sequence alignment. A subsequent report presented statistical deductions of several characteristic measures (Pevzner, 1992) such as the distance expectation and variance for *L-tuple* comparison. The same report

proposes filtration methods based in a prescreening with these metrics. Accordingly, it is possible to filter out sequences with low similarity, those that do not share similar word composition, in order to speed database search for similar sequences. In that report, the same theoretical endpoint proposed previously (Blaisdell, 1989a) is reached. It is noteworthy that these filtration methods are currently being increasingly explored to optimize database search in the face of exponential growth of the sequence repository (URL, 2002).

The statistical properties of Euclidean type distance for *L-tuple* frequencies have been documented further in depth eventually leading to the identification of tests for the non-uniformity of the corresponding distribution based on the Π-statistic thereby defined (Zharkikh and Rzhetsky, 1993). This work enabled the comparison of values obtained for different resolutions and also offers the very interesting promise of a formal link to the determination of evolutionary distances backed by a rate of unit substitution that is not affected by shuffling of conserved segments. The same authors also document a relation between *L-tuple* metric and mismatch count distance, which is the basis for homology estimation by alignment-based methods, thus establishing some comparison between both methods. The validity of those theoretical propositions was accessed in another report with applications to Eubacteria, mitochondria and chloroplasts DNA, including the study of *L-tuple* frequency homogeneity in coding and non-coding regions (Sitnikova and Zharkikh, 1993). The scale dependency of similarity measures itself, such as how 3-*tuple* counts depends on 2-*tuple* counts described in the latter report, is also becoming a recurring theme, albeit reinforced by similar emphasis in the search for unifying scale independent relationships in other areas of biology (Gisiger, 2001).

Weighted Euclidean distance and efficient computation

The fact that the frequency of different words may have a different impact on the standard Euclidean distance between specific words has been explored in the literature to derive weighted measures. The earliest work calculated the weights of individual *L-tuples* in order to maximize the variance of reference sequences with regard to random sequences (Torney *et al.*, 1990). This approach maximizes the discrimination of reference sequence families. The original implementation, maybe due to its relatively pioneering date of 1990, is curiously based on weighting *L-tuple* counts rather than frequencies (Equation 9), where ρ_i is the weight assigned to the ith word. The weighted distances are then combined by summing the weighted count difference at different resolutions, from l to *u-tuples*.

$$d^2(X, Y) = \sum_{L=l}^{u} \sum_{i=1}^{K} \rho_i (c_{L,i}^X - c_{L,i}^Y)^2 \qquad (9)$$

This metric was designated as ***d2 distance*** and has subsequently been used, in its unweighted form, as a stand-alone high performance sequence comparison technique for database search (Hide *et al.*, 1994). The latter work stands on a category of its own due to its focus on heuristic optimization of the computational implementation. That report in particular was directed to the identification of optimum values for word length L, window size and extent of overlap. For the particular example discussed in that report, search for lipases in a genomic database, an optimal resolution of $L = 8$ was found to achieve results similar to performing the search using FASTA.

The practical use of d^2 distance has a published record that continues to the present day including the clustering of EST sequences with full-length cDNA data (Burke *et al.*, 1999) and the recent estimation of the number of human genes (Davison and Burke, 2001). The method has proven to be selective, sensitive and amenable to high performance implementation. These properties, combined with the advantages shared by other alignment-free methods of being context-independent, and consequently the fact that homologous sequences that are scrambled or contain insertions and deletions will still yield a small d^2 value, has had this measure selected for inclusion in software packages. In particular, d^2 clustering was incorporated in the software package STACK (Sequence Tag Alignment and Consensus Knowledgebase), a sequence analysis tool where clustering does not rely on pair-wise alignment (Burke *et al.*, 1998; Christoffels *et al.*, 2001; Hide *et al.*, 1997; Miller *et al.*, 1999). Even more recently, this algorithm was optimized by parallelization (Carpenter *et al.*, 2002), furthering their efficient computation, with a visible relevance for the classification of EST sequences.

In general, it is interesting to note that, very recently, filtration methods based on distance between frequencies of words have had their usage greatly increased as procedures to 'seed' a conventional alignment, both for DNA sequences (Giladi *et al.*, 2002) and for proteins (Coghlan *et al.*, 2001). Both FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990) rely on seeding for a preselection of candidate sequences for alignment. Indeed, pre-processing sequence querying by efficient elimination of non-similar candidates appears to be the path through which alignment-free sequence comparison is gradually being incorporated in widely used bioinformatics applications.

CORRELATION STRUCTURE

Once the conversion of sequences into *L-tuple* frequencies was established, a variety of metric systems were quickly proposed, as described above for Euclidean distances. Within this context, the proposition of metric distances

between sequences based on the correlation coefficients was to be expected (Fichant and Gautier, 1987; Gibbs *et al.*, 1971; van Heel, 1991). Indeed, that approach has since been put to practice to classify proteins based on di-peptide frequencies (Petrilli, 1993). The calculation of the linear correlation coefficient (LCC) between two sequences X and Y, from *L-tuple* frequencies, f_L^X and f_L^Y, uses the conventional Pearson formalism as detailed in Equation (10).

$$d_L^{\mathrm{LCC}}(X,Y) = \left[K\sum_{i=1}^{K} f_{L,i}^X \cdot f_{L,i}^Y - \sum_{i=1}^{K} f_{L,i}^X \cdot \sum_{i=1}^{K} f_{L,i}^Y\right] \Big/ \left[\left[K\sum_{i=1}^{K}(f_{L,i}^X)^2 - \left(\sum_{i=1}^{K} f_{L,i}^X\right)^2\right]^{1/2} \times \left[K\sum_{i=1}^{K}(f_{L,i}^Y)^2 - \left(\sum_{i=1}^{K} f_{L,i}^Y\right)^2\right]^{1/2}\right] \tag{10}$$

This can be simply expressed by taking vectors f_L^X and f_L^Y as pairs in $\Re^2$, by plotting the K points $\left(f_{2,i}^X, f_{2,i}^Y\right)$, and calculating the correlation coefficient R.

As noted before for Euclidean distances, the availability of a correlation based, alignment-free, sequence comparison method is of immediate advantageous use for the querying of large sequence databases, and has been applied to protein database searching (Petrilli and Tonukari, 1997). The applied work yielded a number of simplifying conclusions that greatly enhance its practical value, such as the fact that only 25 out of 400 possible dipeptide frequencies were needed to correctly classify protein families (Solovyev and Makarova, 1993).

The way tuples are defined has itself been the object of exploration with the goal of identifying spatial correlations between positions differently spaced apart in the sequence (Mironov and Alexandrov, 1988). Although this approach has not been subsequently pursued by other researchers, its original proposition took place in the very early period of development of alignment-free methods and offers a different perspective on the conceptual foundations of this field. The spatial correlation measure is based on the determination of dimeric tuples ($L = 2$) where the first and second positions are separated by a fixed arbitrary number of units. The original report proposed to screen different values for the separation and combination of the results in a single correlation measure. The difference between sequences was then developed using the Euclidean distance of the vectors representing the extracted features.

COVARIANCE METHODS

The methods reviewed above explore the use of Euclidean distances and correlations between *L-tuple* representations of sequences. This section reviews, instead, distances that take into account the data covariance structure. In this context the use of Mahalanobis distances (Equation 11) and standardized Euclidean distances (Equation 12), play a central role.

$$d_L^{\mathrm{M}}(X,Y) = (c_L^X - c_L^Y)^{\mathrm{T}} \cdot S^{-1} \cdot (c_L^X - c_L^Y) = \sum_{i=1}^{K}\sum_{j=1}^{K}(c_{L,i}^X - c_{L,i}^Y) \cdot s_{ij}^{\mathrm{inv}} \cdot (c_{L,j}^X - c_{L,j}^Y) \tag{11}$$

In Equation (11), $\mathbf{S} = [s_{ij}]$ represents the covariance matrix of *L-tuple* counts, which inverted is composed of $K \times K$ elements s_{ij}^{inv}. The standard Euclidean distance (Equation 12) forces $\mathrm{cov}(c_i, c_j) = 0$ for $i \neq j$. Therefore, in this distance measure the correlations between different words are ignored and only same word variances are accounted for.

$$d_L^{\mathrm{SE}}(X,Y) = (c_L^X - c_L^Y)^{\mathrm{T}} \cdot [\mathrm{diag}(s_{11},\ldots,s_{KK})]^{-1} \cdot (c_L^X - c_L^Y) = \sum_{i=1}^{K} \frac{(c_{L,i}^X - c_{L,i}^Y)}{s_{ii}} \tag{12}$$

The relevance of this simplification is put into context by noting that the standard Euclidean distance (Equation 12) is reduced to the squared Euclidean distance (Equation 8) if the variance structure is ignored, i.e. if $s_{ii} = 1$, $i = 1,\ldots,K$. Both the Mahalanobis and standard Euclidean distance were first proposed for sequence comparison relatively recently (Wu *et al.*, 1997). In that report the author also proposes to combine different resolutions to obtain a unique distance measure (Equation 13), similarly to the approach followed in the definition of the $\boldsymbol{d}^2$ measure (Equation 9).

$$d^{M*} = \sum_{L=l}^{n} d_L^{\mathrm{M}}$$
$$d^{SE*} = \sum_{L=l}^{n} d_L^{\mathrm{SE}} \tag{13}$$

It is in the context of these metrics that the measure of overlap capability between words, introduced in the **Background** section above, is most relevant. Overlap capability indicates periodicity in the word, which leads to higher probability of co-occurrence of words sharing the repeated motifs (Gentleman and Mullin, 1989; Reinert *et al.*, 2000), consequently altering the covariance structure presented.

Some implementation problems arise when calculating Mahalanobis distance: the covariance matrix **S** has determinant near zero (matrix almost singular) so it is computationally difficult to calculate its inverse. A solution often proposed that was followed to overcome this problem is to use pseudo-inverse matrices (Wu *et al.*, 1997). However this is unsatisfactory for word lengths higher than 4, when the computational load becomes too heavy for practical implementation. For this reason and although important from a theoretical point of view, this method was ruled out by the proponent for applications with long alphabets and/or long sequences. Nevertheless, it was shown to be very efficient when challenged with finding human lipoprotein lipase (LPL) in a database, providing better selectivity and sensibility than previous distances, namely the Euclidean and standard Euclidean measures.

The Mahalanobis based distance was also proposed for protein classification in a report (Solovyev and Makarova, 1993) already approached in the **Correlation** section. With regard to the Mahalanobis distance, the proponents suggested practical simplifications, namely that only oligopeptides whose frequencies are distinct from random proteins, are used, as these are the most informative and, consequently, the most discriminant data.

INFORMATION THEORY-BASED MEASURES

The methods reviewed above were based on statistical distances between frequency vectors. Instead, the distances reviewed in this section are based on the same *L-tuple* vectors as above but an information theory based metric is used to quantify the dissimilarity between them. To that effect, the Kullback–Leibler discrepancy, KL (see **Background** section), was recently proposed (Wu *et al.*, 2001). The KL discrepancy between sequences X and Y, is computed from their *L-tuple* frequencies (Equation 14).

$$d_L^{KL}(X,Y) = \sum_{i=1}^{K} f_{L,i}^X \cdot \log_2\left(\frac{f_{L,i}^X}{f_{L,i}^Y}\right). \qquad (14)$$

To avoid having an infinite $d_L^K(X,Y)$ when $f_{L,i}^Y = 0$, the authors also suggest modifying this formulation (Equation 14) by adding a unit to both terms of the frequency ratio. As with the Mahalanobis distance, this report also proposes an implementation by sliding partially overlapping windows to select the best conserved regions, under the assumption of contiguity discussed above. The KL distance was validated using the human lipoprotein lipase data set the same authors had previously used to evaluate the use of Mahalanobis distance (Wu *et al.*, 1997). It was concluded that the best performing metric with regard to selectivity and sensitivity was the Mahalanobis distance (Equation 11), followed closely by the standard Euclidean distance (Equation 12) and somewhat further behind by the KL discrepancy (Equation 14). These three distance measures clearly outperformed the conventional Euclidean distance (Equation 8). As regards computational efficiency, the performances are reversed with KL discrepancy (Equation 14) being preferred, followed by the standard Euclidean distance (Equation 13). The Mahalanobis distance, as mentioned above, has a hefty computational cost associated to the calculation of the inverse covariance matrices $\mathbf{S}^{-1}$ (Equation 11).

ANGLE METRICS

Very recently, (Stuart *et al.*, 2002a,b), a new metric was proposed that falls on a category of its own where the distance between two sequences is based on the angle between the *L-tuple* count vectors (Equation 15). As these vectors usually have high dimensionality ($K = r^L$, see Equation 1), single value decomposition (SVD) is applied before calculating the angle cosine. Only the dimensions with the higher eigenvalues are used, thus substantially reducing dimensionality with the additional advantage of filtering some noise from this information. Dimensionality reduction along similar lines has been reported by other authors as being very useful for information retrieval from databases (Berry *et al.*, 1999).

$$d_L^{\cos}(X,Y) = \theta_{XY}, \text{ where } \cos(\theta_{XY}) = \frac{(c_L^X)^{\mathrm{T}} \cdot c_L^Y}{\|c_L^X\| \cdot \|c_L^Y\|} = \frac{\sum_{i=1}^{K} c_{L,i}^X \cdot c_{L,i}^Y}{\sqrt{\sum_{i=1}^{K}(c_{L,i}^X)^2} \cdot \sqrt{\sum_{j=1}^{K}(c_{L,j}^Y)^2}} \qquad (15)$$

Interestingly, this metric is not sensitive to repetitions, instead returning the difference between the motifs. For example, if a sequence X is compared with its double repetition XX, the vectors c of the counts will have different norms but will have the same direction in space, because $c^X = 2c^{XX}$, causing the angle distance between them to be zero. This property is of fundamental value because it automatically filters repetitions, therefore distinguishing sequences by the different balance of tuple composition only. It is also interesting to note that the distance proposed has strong similarities to the correlation distance d_L^{LCC} (Equation 10). The pair-wise cosine values were proposed in the same reports to convert to evolutionary distance, determined from *L-tuple* counts, as detailed in Equation (16) (Stuart *et al.*, 2002a,b).

$$d_L^{\mathrm{EVOL}}(X,Y) = -\ln[(1+\cos\theta_{XY})/2] \qquad (16)$$

The cross-tabulation of Evolutionary distances was then inputted to the NEIGHBOR program (Saitou and Nei, 1987), part of the PHYLIP package (Felsenstein, 1993), used to construct the corresponding phylogenetic trees. The choice of the appropriate L-resolution is further

discussed by the proponent whose results suggest it may be specific to the degree of evolutionary divergence. In particular, d_L^{EVOL} was applied to the study of whole mitochondrial genome, and the resulting evolutionary distances were observed to be in agreement with the values previously obtained by other methods. That work put particular emphasis on the dimensionality reduction using the SVD algorithm, which allows a different and interesting interpretation of this metric: by reducing the basis vectors of the representation, the authors are somehow neglecting the main L-tuple composition used, looking for some feature space that conveys a special non-literal representation, in some sense. This can provide a pattern analysis beyond word composition. In principle, the technique could be equally relevant and applied to the preceding metrics.

Resolution-free methods

The metrics reviewed above are dependent on a specific resolution or word length of the *L-tuples*. This problem was solved in some reports cited above by choosing the best discriminant resolution or combining results obtained with arbitrary word-length intervals. Instead, this section reviews alignment-free sequence comparison methods that do not resolve to fixed word-length distance measures, which represents absolute independence from the assumption of conservation of contiguity. This goal has been pursued following two alternative paths. The first one uses sequence compression as a tool to measure sequence complexity. The extent to which joint compression is more effective than independent compression is used as a measure of similarity. The second approach focuses on the representation of the sequence itself, using iterative functions as bijective maps to continuous, scale-independent formats, where resolution-free comparisons can be pursued.

UNIVERSAL SEQUENCE MAPS (USM)

The pursuit of distance measures independent from *L-tuple* resolution has been proposed by seeking sequence representations that would themselves be scale independent. Chaos Theory, namely as regards the use of iterative functions, is at the foundation of this pursuit. The proposition of iterative functions for the representation of biological sequences is now over a decade old. The original report identified an iterative function for DNA representation, which was named Chaos Game Representation, CGR (Jeffrey, 1990). The recognition that CGR defines a resolution free transition matrix that can be used to derive distance metrics is much more recent (Almeida *et al.*, 2001). That work was later extended and generalized for any order alphabets, thus enabling the study of any discrete sequence, and the new iterative function was renamed Universal Sequence Maps, USM (Almeida and Vinga, 2002). The interesting novel property of the USM bijective mapping is the possibility of accurately representing and summarizing any sequence in a continuous multidimensional space at arbitrary resolution (that can be later used to recover sequence context). The comparison of any two unit positions will yield the level of identity between the respective regions in the sequence. For example, the representation of two symbols $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ in USM coordinates can be used to estimate the difference between those symbols in the original sequence (Equation 17).

$$d^{\mathrm{USM}}(a, b) = -\log_2(\max_i |a_i - b_i|) \qquad (17)$$

The USM method can be applied to DNA, proteins and natural language texts but it still is in experimental development and has not been yet completely tested in challenging sequence sets. It would also be desirable to apply this methodology to multiple comparison and database queries. It should also be noted that the metric proposed, although taking into account symbol context, does not define an overall sequence dissimilarity, like previously reviewed distances.

KOLMOGOROV COMPLEXITY

The use of savings in joint compression as a measure of similarity is founded on information theory and coding, particularly on Kolmogorov complexity theory. Similarly to the methods reviewed in the last two sections, this one is also a very recent proposition (Li *et al.*, 2001). The fundamental concept behind the distance metric proposed is that of algorithmic complexity. In practice, this pursuit requires the use of compression algorithms that are assumed to be efficient. There are presently no absolute measures of algorithmic complexity, which can only be estimated. (For a review of methods see V'Yugin (1999).) In that report (Li *et al.*, 2001), sequence compression is performed using the GenCompress software program (Chen *et al.*, 1999), empirically assessing the Kolmogorov complexity, $K(X)$, of a sequence, X by the length of its compressed representation. The conditional complexity is obtained by compressing the juxtaposition of both sequences. The distance measure derived thereof, d^{KC}, detailed in Equation (18), uses the relative decrease in complexity or conditional complexity $K(X \mid Y)$ as a measure of sequence similarity (Li and Vitanyi, 1997).

$$d^{\mathrm{KC}}(X, Y) = 1 - \frac{K(X) - K(X \mid Y)}{K(XY)} \qquad (18)$$

The authors demonstrate that d^{KC} satisfies the axioms of a distance function (Equation 5). This method was only tested with mammalian complete mitochondrial genomes (mtDNA), and the distances obtained were observed to

be consistent with the known phylogenetic relationships. Despite this method was not yet fully explored, only in a rather limited set of sequences, and the need to estimate the quantities evolved, namely $K(.)$, by a compressing algorithm, it is conceptually attractive and elegant which suggests its further study and extension to higher order alphabets, for example, in comparing proteins.

RECENT EXPLOITS

The increase in diversity of the newer alignment-free distance measures being proposed beyond the framework reviewed here is very apparent as this review is finalized. For example, alignment-independent classification of G-protein coupled receptors (GPCR) based in extracting physical properties of amino acids has been very recently suggested (Lapinsh *et al.*, 2002). This correlation data was processed with multivariate statistical methods, namely principal component analysis (PCA), partial least squares (PLS), autocross-covariance transformations (ACC's), *z*-scores, in order to weight the individual properties as to correctly classify the proteins studied in super-families. Previous attempts to GPCR classification without alignment were based on the extraction of statistics of communality and specificity for each *L-tuple* (Daeyaert *et al.*, 1998). These characteristics measure the relative frequency of specific words with regard to the respective super-families.

ALGORITHM IMPLEMENTATION/TOOLBOX PRESENTATION

Most of the distance metrics reviewed in this report were coded anew and tested. For that purpose a software toolbox was written in MATLAB language and is made publicly available by the reviewers at http://www.bioinformatics.musc.edu/resources.html.
Submission of new distance metrics or more efficient implementation of existing ones to that web-based repository is encouraged. The toolbox includes a small manual that explains the algorithms and the use of the functions. It also includes a set of test sequences using different alphabets. Three data sets are included (submission of particularly challenging sets is also encouraged): closely related DNA sequences for which alignment is still a good solution, related protein sequences for which only moderate or weak alignments can be produced, and natural languages—the same text in ten western European idioms with clearly recognizable philology.

CONCLUSIONS

Sequence comparison by alignment has both fundamental and computational limitations. The conservation of contiguity underlying alignment is at odds with genetic recombination, which includes shuffling subgenomic DNA fragments. This limitation is particularly clear by recalling that, regardless of the progress in the identification of scoring matrices, alignment fails to recognize proteomic sequences with less than 20% sequence identity. In addition, optimal alignment is computationally too heavy for efficiently querying the sharply inflating genomic and proteomic public databases. The increasing awareness of those limitations is driving the proposition of a diversity of new foundations for alignment-free sequence analysis, hereby reviewed. The diversity of theoretical foundations explored by the reports reviewed here ranges from Linear Algebra and Statistics, to Information Theory, Kolmogorov complexity and Chaos Theory. The recent abundance of successful applications of alignment-free sequence analysis, and the increasing focus on practical implementations makes it a safe prediction that the next few years will see some of them become widely used for functional annotation and phylogenetic study.

ACKNOWLEDGEMENTS

The authors thankfully acknowledge the financial support by grants SFRH/BD/3134/2000 and Sapiens/34794/99 of Fundação para a Ciência e Tecnologia of the Portuguese Ministry of Science and Technology (FCT/MCT).

REFERENCES

Almeida,J.S. and Vinga,S. (2002) Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, **3**, 6.

Almeida,J.S., Carriço,J.A., Maretzek,A., Noble,P.A. and Fletcher,M. (2001) Analysis of genomic sequences by chaos game representation. *Bioinformatics*, **17**, 429–437.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ash,R.B. (1990) *Information Theory*. Dover, New York.

Attwood,T.K. (2000) Genomics: the Babel of bioinformatics. *Science*, **290**, 471–473.

Berry,M.W., Drmac,Z. and Jessup,E.R. (1999) Matrices, vector spaces, and information retrieval. *SIAM Review*, **41**, 335–362.

Blaisdell,B.E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl Acad. Sci. USA*, **83**, 5155–5159.

Blaisdell,B.E. (1989a) Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. *J. Mol. Evol.*, **29**, 538–547.

Blaisdell,B.E. (1989b) Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. *J. Mol. Evol.*, **29**, 526–537.

Burke,J., Wang,H., Hide,W. and Davison,D.B. (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.*, **8**, 276–290.

Burke,J., Davison,D. and Hide,W. (1999) d2_cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.*, **9**, 1135–1142.

Carpenter,J.E., Christoffels,A., Weinbach,Y. and Hide,W.A. (2002) Assessment of the parallelization approach of d2_cluster for high-performance sequence clustering. *J. Comput. Chem.*, **23**, 755–757.

Chen,X., Kwong,S. and Li,M. (1999) A compression algorithm for DNA sequences and its applications in genome comparison. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 51–61.

Christoffels,A., van Gelder,A., Greyling,G., Miller,R., Hide,T. and Hide,W. (2001) STACK: sequence tag alignment and consensus knowledgebase. *Nucleic Acids Res.*, **29**, 234–238.

Coghlan,A., MacDonaill,D.A. and Buttimore,N.H. (2001) Representation of amino acids as five-bit or three-bit patterns for filtering protein databases. *Bioinformatics*, **17**, 676–685.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley, New York.

Daeyaert,F., Moereels,H. and Lewi,P.J. (1998) Classification and identification of proteins by means of common and specific amino acid n-tuples in unaligned sequences. *Comput. Methods Programs Biomed.*, **56**, 221–233.

Davison,D.B. and Burke,J.F. (2001) Brute force estimation of the number of human genes using EST clustering as a measure. *IBM J. Res. Dev.*, **45**, 439–447.

Dayhoff,M.O., Schwartz,R. and Orcutt,B. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of protein sequence and structure*, National Biomedical Research Foundation, Vol. 5, supplement 3, Washington, DC, pp. 345–352.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press.

Everitt,B., Landau,S. and Leese,M. (2001) *Cluster Analysis*. Arnold, London.

Felsenstein,J. (1993) *PHYLIP (Phylogeny Inference Package)*, Department of Genetics, University of Washington, Seattle.

Fichant,G. and Gautier,C. (1987) Statistical method for predicting protein coding regions in nucleic acid sequences. *Comput. Appl. Biosci.*, **3**, 287–295.

Fuchs,R. (2002) From sequence to biology: the impact on bioinformatics. *Bioinformatics*, **18**, 505–506.

Gentleman,J.F. and Mullin,R.C. (1989) The distribution of the frequency of occurrence of nucleotide subsequences, based on their overlap capability. *Biometrics*, **45**, 35–52.

Gibbs,A.J., Dale,M.B., Kinns,H.R. and MacKenzie,H.G. (1971) The transition matrix method for comparing sequences; its use in describing and classifying proteins by their amino acids sequence. *Systematic Zool.*, **20**, 417–425.

Giladi,E.G., Walker,M., Wang,J.Z. and Volmuth,W. (2002) SST: an algorithm for finding near-exact sequence matches in time proportional to the logarithm of the database size. *Bioinformatics*, **18**, 873–877.

Gisiger,T. (2001) Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biol. Rev. Camb. Philos. Soc.*, **76**, 161–209.

Gordon,A.D. (1999) *Classification*, Chapman & Hall/CRC, Boca Raton, FL.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

van Heel,M. (1991) A new family of powerful multivariate statistical sequence analysis techniques. *J. Mol. Biol.*, **220**, 877–887.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hide,W., Burke,J. and Davison,D.B. (1994) Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J. Comput. Biol.*, **1**, 199–215.

Hide,W., Burke,J., Christoffels,A. and Miller,R. (1997) A novel approach towards a comprehensive consensus representation of the expressed human genome. In Miyano,S.a.T.,T. (ed.), *Genome Informatics*. Universal Academy Press, Tokyo, Japan, pp. 187–196.

Jeffrey,H.J. (1990) Chaos game representation of gene structure. *Nucleid Acids Res.*, **18**, 2163–2170.

Koonin,E.V. (1999) The emerging paradigm and open problems in comparative genomics. *Bioinformatics*, **15**, 265–266.

Kullback,S. (1968) *Information theory and statistics*. Dover, New York.

Lapinsh,M., Gutcaits,A., Prusis,P., Post,C., Lundstedt,T. and Wikberg,J.E. (2002) Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci.*, **11**, 795–805.

Li,M. and Vitanyi,P. (1997) *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York.

Li,M., Badger,J.H., Chen,X., Kwong,S., Kearney,P. and Zhang,H. (2001) An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, **17**, 149–154.

Lynch,M. (2002) Intron evolution as a population-genetic process. *Proc. Natl Acad. Sci. USA*, **99**, 6118–6123.

Miller,R.T., Christoffels,A.G., Gopalakrishnan,C., Burke,J., Ptitsyn,A.A., Broveak,T.R. and Hide,W.A. (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.*, **9**, 1143–1155.

Mironov,A.A. and Alexandrov,N.N. (1988) Statistical method for rapid homology search. *Nucleic Acids Res.*, **16**, 5169–5173.

Needleman,S.B. and 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.

Pearson,W.R. (2000) Protein sequence comparison and protein evolution. *Tutorial—ISMB2000*.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Petrilli,P. (1993) Classification of protein sequences by their dipeptide composition. *Comput. Appl. Biosci.*, **9**, 205–209.

Petrilli,P. and Tonukari,N.J. (1997) PFDB: a protein families database for Macintosh computers. The effectiveness of its organization in searching for protein similarity. *J. Protein Chem.*, **16**, 713–720.

Pevzner,P.A. (1992) Statistical distance between texts and filtration methods in sequence comparison. *Comput. Appl. Biosci.*, **8**, 121–127.

Régnier,M. (1998) A unified approach to word statistics. In Press,A. (ed.), *Proceedings of the Second Annual International Conference on Computational Molecular Biology*. New York, pp. 207–213.

Reichhardt,T. (1999) It's sink or swim as a tidal wave of data approaches. *Nature*, **399**, 517–520.

Reinert,G., Schbath,S. and Waterman,M.S. (2000) Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.*, **7**, 1–46.

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

Schott,J.R. (1997) *Matrix Analysis for Statistics*. Wiley, New York.

Searls,D.B. (2001) Reading the book of life. *Bioinformatics*, **17**, 579–580.

Shannon,C.E. (1948) A mathematical theory of communication. *The Bell System Technical J.*, **27**, 379–423, 623–656.

Siegmund,D. and Yakir,B. (2000) Approximate p-values for local sequence alignments. *The Annals of Statistics*, **28**, 657–680.

Sitnikova,T.L. and Zharkikh,A.A. (1993) Statistical analysis of L-tuple frequencies in eubacteria and organelles. *Biosystems*, **30**, 113–135.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Solovyev,V.V. and Makarova,K.S. (1993) A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. *Comput. Appl. Biosci.*, **9**, 17–24.

Storey,J.D. and Siegmund,D. (2001) Approximate *p*-values for local sequence alignments: numerical studies. *J. Comput. Biol.*, **8**, 549–556.

Strang,G. (1988) *Linear Algebra and Its Applications*. Thomson, London.

Stuart,G.W., Moffett,K. and Baker,S. (2002a) Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, **18**, 100–108.

Stuart,G.W., Moffett,K. and Leader,J.J. (2002b) A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol. Biol. Evol.*, **19**, 554–562.

Torney,D.C., Burks,C., Davison,D. and Sirotkin,K.M. (1990) Computation of d2: a measure of sequence dissimilarity. In George,I. and Bell,T.G.M. (eds), *Computers and DNA : the proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop, held December 12 to 16, 1988 in Santa Fe, New Mexico*. Addison-Wesley, Redwood City, CA, pp. 109–125.

URL (2002) http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

V'Yugin (1999) Algorithmic complexity and stochastic properties of finite binary sequences. *The Computer J.*, **42**, 294–317.

Waterman,M.S. (1995) *Introduction to Computational Biology: Maps, Sequences, and Genomes: Interdisciplinary Statistics*. Chapman and Hall/CRC, Boca Raton, FL.

Wu,T.J., Burke,J.P. and Davison,D.B. (1997) A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, **53**, 1431–1439.

Wu,T.J., Hsieh,Y.C. and Li,L.A. (2001) Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, **57**, 441–443.

Zhang,Y.X., Perry,K., Vinci,V.A., Powell,K., Stemmer,W.P. and del Cardayre,S.B. (2002) Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature*, **415**, 644–646.

Zharkikh,A.A. and Rzhetsky,A. (1993) Quick assessment of similarity of two sequences by comparison of their L-tuple frequencies. *Biosystems*, **30**, 93–111.