BIOINFORMATICS

Vol. 18 Suppl. 2 2002
Pages S44–S53

A new distance measure for comparing sequence profiles based on path lengths along an entropy surface

Gary Benson

Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, USA

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

We describe a new distance measure for comparing DNA sequence profiles. For this measure, columns in a multiple alignment are treated as character frequency vectors (sum of the frequencies equal to one). The distance between two vectors is based on minimum path length along an entropy surface. Path length is estimated using a random graph generated on the entropy surface and Dijkstra's algorithm for all shortest paths to a source. We use the new distance measure to analyze similarities within familes of tandem repeats in the *C. elegans* genome and show that this new measure gives more accurate refinement of family relationships than a method based on comparing consensus sequences.

INTRODUCTION

In this paper, we describe a new distance measure for comparing sequence profiles and then apply it to the comparison of tandem repeats in DNA. A *tandem repeat* is an occurrence of two or more adjacent, often *approximate* copies of a sequence of nucleotides. Tandem repeats are ubiquitous sequence features in both prokaryotic and eukaryotic genomes. In humans, they are known to cause at least ten inherited neurological diseases including fragile-X mental retardation (Verkerk *et al.*, 1991), Huntington's disease (Huntington's disease collaborative research group, 1993), and myotonic dystrophy (Fu *et al.*, 1992) and they are associated with a number of other major diseases, including diabetes (Owerbach and Gabbay, 1993; Bennett *et al.*, 1995), epilepsy (Virtaneva *et al.*, 1997; Lalioti *et al.*, 1997), and certain cancers (Phelan *et al.*, 1996). Tandem repeats are used for DNA fingerprinting and have recently been used to discriminate between different bacterial strains, including anthrax strains (Keim *et al.*, 2000; Flèche *et al.*, 2001).

Although tandem repeats form one of the major classes of repeats in genomic DNA and despite their biological importance, their detailed study as a class is only just beginning. An important initial step is the grouping of repeats into *families* in order to identify and study their common properties. Members of a family have similar sequence but occur at different locations in a genome or in different genomes. Families have been detected in both prokaryotic and eukaryotic genomes, including the *E. coli*, *P. aeruginosa*, *S. cerevisiae*, *C. elegans*, and human genomes.

We are interested in clustering repeats into families based on their sequence similarities. Clustering of high dimensional objects such as sequences can often be effectively accomplished using a table of inter-object distances. The complexity of creating such a table for tandem repeats stems from the variations that are commonly observed in families. For example, related repeats often differ in copy number, display mutational differences between some copies and not others and exhibit shuffled orderings of the mutationally different copies, one repeat to another (see Figure 1 for examples). Such variations have been used by Kececioglu and Yu (2001) to separate repeats incorrectly overlayed during sequence assembly. As a consequence of these differences, comparison by standard methods such as BLAST (Altschul *et al.*, 1990) or alignment (Smith and Waterman, 1981; Benson, 1997) is problematic because they will tend to focus on the wrong properties, such as copy number, ordering, or strong similarity between some of the copies.

An accurate and effective comparison scheme will be insensitive to these properties. One promising approach is to represent each repeat by a *profile* (Gribskov *et al.*, 1990) of its copies and then measure the similarity of the profiles with a cyclic alignment algorithm (Maes, 1990; Benson, 2000). A profile is a sequence whose length equals the number of columns in a multiple alignment and whose individual elements are the *character compositions* of the columns. If we assume that the ordering of the characters in a column is not informative (due, for example, to shuffling), then each composition is a vector of character frequencies. Since these frequencies sum to

one, each composition can also be thought of as a *discrete probability distribution*.

Alignment of profiles requires a distance function for composition pairs. Such functions, in the language of probability distributions, are called *divergence measures*. For our application, the function should return intuitively appropriate distances in three common situations: (1) a large distance when the characters contained in one composition are not contained in the other, (2) a smaller distance when the dominant character in both compositions is the same, but the minor characters differ, and (3) a small distance when compositions share the same characters but differ in their ratios or which is dominant. (See Figure 1 arrows.) Common ancestry presumably ranks the rarity of these situations in the order given because it would require (1) loss and/or replacement of all nucleotides at the same location in every copy, probably through point mutation followed by duplication and excision of copies, (2) loss or replacement of only a few nucleotides by similar methods, and (3) change in ratio of nucleotides probably by duplication and excision alone.

A variety of *entropy functions* have been used as divergence measures. Examples are the relative entropy and the symmetric relative entropy (Kullback, 1968; Lin, 1991). An attractive feature of entropy functions is the initial rapid growth from zero as the composition goes from totally conserved (all the same character) to less conserved. Unfortunately, these functions are undefined (divide-by-zero) when one distribution has a character that the other lacks. Alternative functions include the K, L and Jensen-Shannon divergence measures defined by Lin (1991) and similar measures defined by Wong *et al.* (1993). These are relative entropy measures which compute similarity to an *average* or *weighted average* distribution and thereby avoid the divide-by-zero problem. A variation on the Jensen-Shannon measure, used by Yona and Levitt for protein sequences (Yona and Levitt, 2002), weights the distance between the compositions by the distance of their average to a background composition. These functions are extremely sensitive to the location of the average distribution relative to those being tested, do not rank distances as described in the common situations above and produce a range of distance values which span several orders of magnitude making them unsuitable for use in alignment algorithms.

A function commonly used in multiple alignment algorithms is the *sum-of-pairs*, based on the product of matching and mismatching letter frequencies. Related measures have been used by Edwards and Cavalli-Sforza (1964) and Nei *et al.* (1983). Lyngso *et al.* (1999) use the product of matching letter frequencies only in a metric for comparing hidden Markov models. In the case of DNA, where nucleotide bases are generally scored as either a match or a mismatch, sum-of-pairs inappropriately overemphasizes the mismatches and does not give consistent scores when distributions are identical (for example if both distributions are 100% 'A', the sum-of-pairs score is much better than if both distributions are 50% 'A' and 50% 'C').

Below, we describe a new divergence measure based on the shortest path between points along an entropy surface. Its advantages are that it incorporates the rapid growth from zero of entropy functions, avoids divide-by-zero problems, does not base the comparison on an average distribution and appropriately ranks the situations described above. The paper is organized as follows. First, we give formal definitions of profiles and compositions, then we describe the entropy surface, the concept of distance along the surface and some variations. Next, we describe how to estimate path lengths using Dijkstra's algorithm and how we create a distance table. Finally, we describe application of this new distance measure to the analysis of tandem repeat families in the *C.elegans* genome.

PROFILES AND COMPOSITIONS

Assume that we are given a multiple alignment, M, of a set of sequences. M has n rows and k columns. In the case of a *single* tandem repeat sequence, the alignment consists of the n individual copies with the ith row of M containing the ith copy (left-to-right) in the tandem repeat. We let $M_{i,j}$ represents the element in the ith row and jth column of M. Each $M_{i,j}$ contains one of the alphabet symbols from $\Sigma = \{A, C, G, T, -\}$ where $-$ indicates a gap in the alignment. $M_{*,j}$ represent the characters in the jth column. The length of the individual columns need not be the same. For the special case where the tandem repeat contains less than a whole number of copies (i.e. less than n but more than $n-1$, a common occurrence), then the partial copy is in the last row of M, the initial columns contain n letters and the final columns contain $n-1$ letters. $|M_{*,j}|$ is the length of column j, F_σ^j is the number of characters σ occurring in column j and $\sum_\sigma F_\sigma^j = |M_{*,j}|$. A *profile* for M is a sequence $S = C_1 C_2 \cdots C_k$ of *compositions*, C_j, each a vector of frequencies of the characters in $M_{*,j}$: $C_j = (f_A, f_C, f_G, f_T, f_-)$ such that, $\forall\sigma \in \{A, C, G, T, -\}$, $f_\sigma = F_\sigma^j / |M_{*,j}| \leq 1$, and $\sum_\sigma f_\sigma = 1$. We denote a special group of compositions the *conserved compositions*, C_A, C_C, C_G, C_T, C_-, where C_σ is the composition with $f_\sigma = 1.0$ and all other frequencies equal to zero. Later in this paper, when discussing *standard compositions*, we will make use of a different formulation for compositions which we call *composition-by-count*. A composition-by-count is a vector, $CC = (F_A, F_C, F_G, F_T, F_-)$ of the counts for each letter σ. The set of points in $R^{|\Sigma|}$ for which the individual coordinates sum to one is termed the *space of valid compositions*.

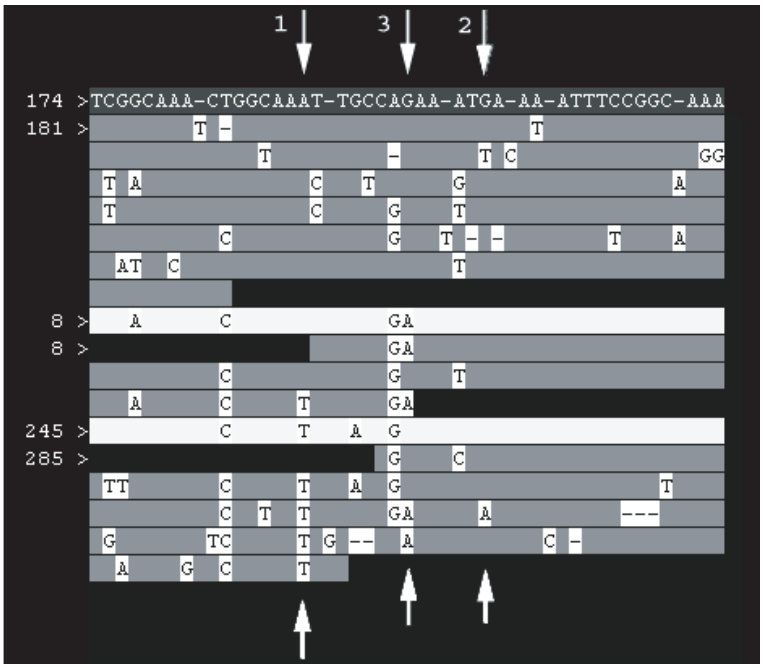


Fig. 1. Three related tandem repeats from *C. elegans*. All sequences are aligned to the consensus pattern (174 - top line) of the first repeat. Light gray lines are the repeats (181, 8, 285), one copy per line. White lines (8, 245) are the remaining consensus patterns. Only differences with consensus 174 are shown. Black indicates where the repeats start and end. Note substitution and indel differences among the repeats, arrowed and numbered as in text: 1) characters in bottom column not contained in top column, 2) majority character the same in both top and bottom columns, minor characters differ, 3) ratio of characters in middle and bottom column differ, majority character is switched.

THE ENTROPY SURFACE

The entropy surface is based on the entropy function

$$H(C) = -\sum_{\sigma} f_\sigma \log(f_\sigma)$$

defined over all possible compositions C (all logarithms used in this paper refer to natural logarithm). In the application here, it is a six dimensional surface (five for the character frequencies and one for the entropy). Figure 2 gives two views of the entropy surface for three character frequencies. The planar triangular region represents the space of valid compositions. The entropy dimension, in actuality a fourth dimension, is plotted perpendicular to the triangular surface. This figure shows the essential features of the entropy surface: (1) the entropy is zero at the conserved compositions (C_A, C_C, etc.) and positive at all other compositions, (2) the steepest rise in entropy occurs at the conserved compositions (the corners) and along the edges of the space of valid compositions (in fact, the slope at these points is infinite), (3) the largest entropy (the central peak) is found at the least conserved composition ($f_A = f_C = f_G$), (4) the surface has 3-fold rotational symmetry (for three character frequencies, 5-fold for five character frequencies).

Distance along the entropy surface

Given two compositions, C_1 and C_2, each defines a point, $P_1 = H(C_1)$ and $P_2 = H(C_2)$ respectively, on the entropy surface. We define the distance, $d(C_1, C_2)$, between these compositions as the shortest path (geodesic) along the entropy surface between P_1 and P_2. It follows that this distance has the three properties of a metric: (1) $d(C_1, C_2) = 0$ iff $C_1 = C_2$, (2) $d(C_1, C_2) = d(C_2, C_1)$, and (3) $d(C_1, C_2) \leq d(C_1, C_3) + d(C_3, C_2)$ (triangle inequality).

Surface variations

As mentioned above, the surface is symmetric. This means that for a five character alphabet, any composition pair (C_1, C_2), belongs to a set of up to 120 composition pairs (5 factorial) whose shortest paths differs only by rotation/mirror imaging of the surface. We call such a set an equivalence class. Each pair in an equivalence class is obtained by applying a particular order permutation to the frequencies in both C_1 and C_2. More formally, let π be a function which maps one 5-tuple into another by permuting the order of the terms and let Π be the set of all such functions π. Then the equivalence class for the

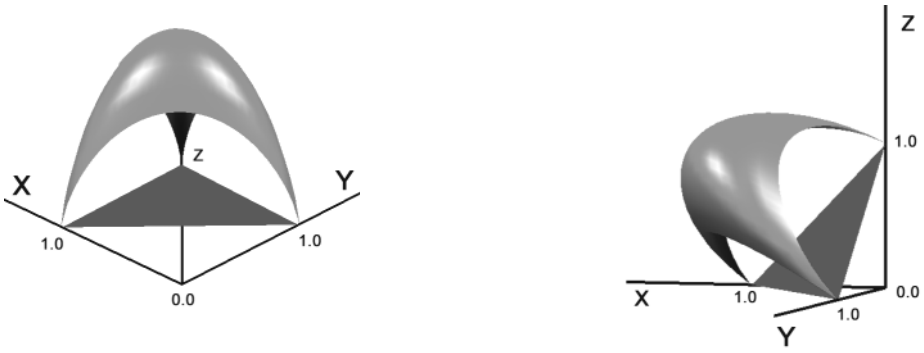


Fig. 2. Two views of the entropy surface for three character frequencies. The surface is drawn perpendicular to the triangular space of valid compositions.

pair (C_1, C_2) is the set

$$\{(C_1', C_2') | \pi \in \Pi \text{ and } C_1' = \pi(C_1) \text{ and } C_2' = \pi(C_2)\}.$$

Equivalence classes with less than 120 pairs occur, for example, when C_1 and C_2 are related by the same permutation function ($C_1 = \pi(C_2)$, $C_2 = \pi^{-1}(C_1)$) or a pair of frequencies is equal in C_1 and the same pair is equal in C_2.

The redundancy caused by symmetry is useful for averaging distances obtained by approximation (see below), but is undesireable if two pairs in the same equivalence class should be treated differently. For example it may be required that the alignment of a column of A's with a column of T's (C_A vs C_T), be scored differently than the alignment of a column of A's with a column of dashes (C_A versus C_-) or the alignment of a column of A's with a column of G's (C_A vs C_G) even though all these pairs occur in the same equivalence class.

The entropy surface can be modified in ways to eliminate the symmetry and reflect the types of differences described above. One modification is to multiply each term in the entropy by its own constant, transforming the entropy function into the following form:

$$H(C) = -\sum_{\sigma} c_\sigma \, f_\sigma \log(f_\sigma).$$

For example, if $c_- = 2$ and the remaining constants equal 1, then the 'peak' of the entropy function is shifted in the direction of 100% dashes. This increases the cost mainly between any composition C_1 and a composition C_2 with a greater frequency of dashes ($f_-^2 > f_-^1$). Figure 3 (left) shows an example of this type of modification for three characters frequencies.

A second possible modification is to include terms of the form $c_{\sigma_1\sigma_2} f_{\sigma_1} f_{\sigma_2}$:

$$H(C) = -\sum_{\sigma} c_\sigma \, f_\sigma \log(f_\sigma) - \sum_{\sigma_1<\sigma_2} c_{\sigma_1\sigma_2} f_{\sigma_1} f_{\sigma_2},$$

with $\sigma_1 < \sigma_2$ determined by lexicographic ordering of the characters. Each new term increases the overall entropy when both f_{σ_1} and $f_{\sigma_2} \neq 0$ and affects path lengths mainly between compositions where the two frequencies do not remain equal. Figure 3 (right) shows an example of this type of modification for three character frequencies.

Variants of the entropy function are not pursued further in this abstract. For the remainder of the paper, we use the symmetric form of the entropy function.

ESTIMATING PATH LENGTHS WITH DIJKSTRA'S ALGORITHM

Estimating path lengths along the entropy surface can be accomplished in different ways. One is to use differential geometry to approximate the geodesic (shortest path). That method is the subject of a separate paper (Ahlbrandt *et al.*, 2000). Here, we use Dijkstra's algorithm for all shortest paths to a source in a weighted graph (Dijkstra, 1959). The graph is a web of connected points placed on the entropy surface. The edges are chords between the points with weight equal to the chord length. Below we describe the construction of this graph and its characteristics.

Standard compositions

Our first approach to building the graph of points involved a set of *standard compositions*. A standard composition is one of the compositions which can occur when every f_σ must be a multiple of $1/N$ for some integer N. For example, we used $N = 30$ so that every frequency had to be a multiple of $1/30$. Another way to look at

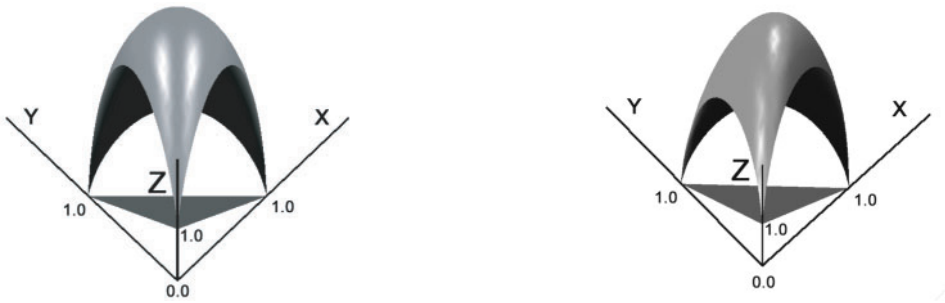


Fig. 3. Two possible variations of the surface. (Left) Near leg taller than remaining two by making $c_\sigma = 2.0$ (Darth Vader look). (Right) All legs asymmetric (note the different arch heights). Produced by making all three $c_{\sigma_1\sigma_2}$ values different.

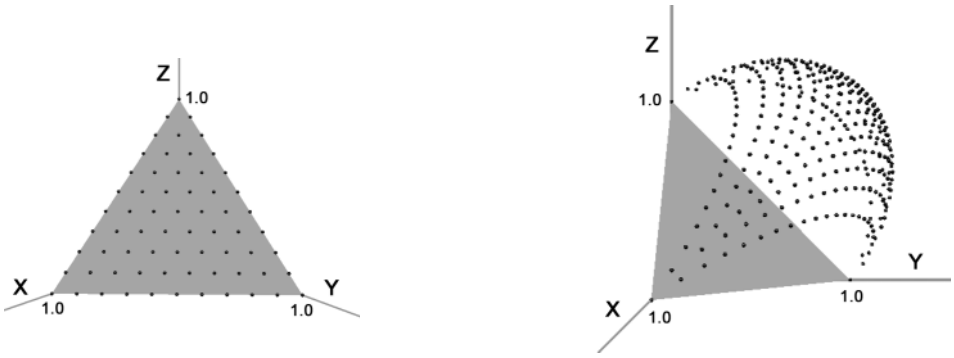


Fig. 4. (Left) Standard composition points on the triangular space of valid compositions for $N = 10$. (Right) Standard composition points projected onto the entropy surface for $N = 23$. Note the larger separation of points between lines versus within lines.

this is to assume that every standard composition has a composition-by-count where the counts F_σ sum to N. For an alphabet of k letters and a column of N characters, the number of different compositions is $\mathbf{C}(N+k-1, k-1)$, where $\mathbf{C}(x, y)$ stands for the number of combinations of x distinct items chosen y at a time. Each standard composition can be given a unique integer identifier. For the five letter alphabet, we use the following formula for the composition-by-count $CC = (F_A, F_C, F_G, F_T, F_-)$: $\mathbf{C}(N+4, 4) - \mathbf{C}(N+3-F_A, 4) - \mathbf{C}(N+2-F_{AC}, 3) - \mathbf{C}(N+1-F_{ACG}, 2) - \mathbf{C}(N-F_{ACGT}, 1)$ where $F_{AC} = F_A + F_C$, etc. A general formula for an alphabet of k letters and a column of N characters is

$$\mathbf{C}(N+k-1, k-1) - \sum_{i=1}^{k-1} \mathbf{C}(N+k-1-i-F_i, k-i)$$

where F_i is the sum of the first i character counts in the composition-by-count vector. Standard composition points form a regular grid in the space of valid compositions. Figure 4 (left) shows this grid for a three letter alphabet ($k = 3$) and $N = 10$.

Our graph based on standard composition points has edges which connect every composition with its *immediate neighbors*. For a composition-by-count, an immediate neighbor is a composition in which one count is reduced and one count is increased by one. Formally, the neighbor set for $CC = (F_A, F_C, F_G, F_T, F_-)$ is $\{(F'_A, F'_C, F'_G, F'_T, F'_-) | \exists \sigma_1, \sigma_2$ with $F'_{\sigma_1} = F_{\sigma_1} + 1, F'_{\sigma_2} = F_{\sigma_2} - 1,$ and $\forall \sigma \notin \{\sigma_1, \sigma_2\}, F'_\sigma = F_\sigma\}$. Visualization of shortest paths obtained with this graph for three characters revealed kinks rather than smooth curves. While standard composition points occur uniformly in the space of valid compositions, they do not occur uniformly on the entropy surface. 'Lines' of points (points in which one of the three counts is held constant) are evident in Figure 4 (right). The cost of jumping from one line to another is so high, that it is cheaper overall to pay the cost once with a kinked path rather than the multiple times required to approximate a smooth curve. Increasing the number of points does not remove the kinks.

Random points

The second approach, which has proven successful, involves the use of *random* points arranged on the entropy surface. We generate the random points uniformly in the space of valid compositions by the method of G. Turk (Turk, 1990) for generating random points in triangles and higher dimensional shapes. Each point generated is a vector of five non-negative frequencies which sum to 1. For an alphabet of five characters, we have had good success with 500 000 random points.

Once generated, the random points are mapped onto the entropy surface and edges connecting them are determined. Two points are connected by an edge if the chord distance between them is less than a minimum cut-off value (below). The chord distance, d_c, for points with composition $C_1 = (f_A^1, f_C^1, f_G^1, f_T^1, f_-^1)$ and $C_2 = (f_A^2, f_C^2, f_G^2, f_T^2, f_-^2)$ is defined as the Euclidian distance in six dimensions between the points on the entropy surface:

$$d_c = \sqrt{\left(\sum_\sigma (\Delta f_\sigma)^2 + (\Delta H)^2\right)}$$

where $\Delta f_\sigma = f_\sigma^1 - f_\sigma^2$ and $\Delta H = H(C_1) - H(C_2)$.

The graph is constructed point by point and for each new point v, it is necessary to test existing points to find those which must be connected to v. We limit the number of existing points tested by establishing a dense set of *regional buckets* which hold the points. Each bucket is defined by a standard composition point (not a true point in the graph). For this purpose we have used $N = 20$ which gives 10 626 buckets. Each new point v is assigned to the bucket whose standard point is closest as determined by the chord distance defined above. Then v tests for connection *only* with the points in neighboring buckets, i.e. those bounding the five dimensional 'cell' containing v. To find neighboring buckets, we determine, for each frequency, f_σ, in v, the two integer values which bracket f_σ in a standard composition. For example if $f_\sigma = 0.37$ and $N = 20$, the two integer values $floor(f_\sigma N)$ and $ceiling(f_\sigma N)$ are 7 and 8. Each combination of the bracketing values for the five frequencies that sum to N defines a neighboring standard composition.

The cut-off distance for connecting points is the Euclidian distance between neighbor standard composition points in the space of valid compositions (ignoring entropy). This corresponds to the closest distance between standard points in the flat 'center' of the entropy surface where ΔH is essentially zero. Two neighboring standard composition points are separated by a change of $1/N$ in two frequencies, yielding a Euclidian distance of $\sqrt{(2/N^2)}$. For $N = 20$ the cutoff distance is 0.07.

As mentioned before, points generated uniformly in the space of valid composition will not be uniform on the entropy surface. The points are most dense in the 'center' and least dense at the 'corners' (all but one f_σ approximately equal to zero) and 'edges' (all but two, three or four f_σ approximately equal to zero). We make two adaptations to mitigate these disparities. First, in order to limit the number of edges in the 'center' of the graph, each bucket is allowed a maximum capacity of 50 points. Second, in order to guarantee that the corners and edges are connected to the remainder of the graph, we generate extra points in these regions: (1) 600 extra points in each corner (a typical corner region is defined by the composition (1, 0, 0, 0, 0), and four additional points obtained by permuting the four final frequencies in $(1 - n, n, 0, 0, 0)$ for $n = 0.15$), (2) 100 extra points on each 2D edge (three $f_\sigma = 0$), and (3) 2000 extra points on each 3D edge (two $f_\sigma = 0$). These values were arrived at by trial and error.

In addition to the random points, 1001 standard composition points ($k = 5, N = 10$) were added to the graph and used to define the distance table (below). Finally, our graph of 501 001 points, generated approximately 23 million edges and took approximately 4 minutes to construct.

CREATING THE DISTANCE TABLE

Our ultimate goal is to compare repeats both within and across genomes. In the case of *C. elegans* alone, this entails alignments that involve over 4000 unique compositions. Rather than compute distances dynamically for such a large set, we have chosen to precompute a modest size distance table for standard compositions and use this table when comparing profiles (next section). In order to create a distance table of reasonable size, we have utilized a standard composition grid ($k = 5, N = 10$) of 1001 *reference* points on the entropy surface and calculated the distance between every pair of these points (roughly 1 million pairs). A standard heap implementation (Cormen *et al.*, 1990) of Dijkstra's single source, all destinations algorithm (Dijkstra, 1959) was used to find the distance from a single reference point to all other reference points, repeated once for every reference point as the source. This procedure, the most time consuming, took approximately 6.3 hours. Figure 5 shows the shortest estimated path between two points on the surface for three character frequencies. We make two modifications to the raw distances to obtain our final distance table.

Averaging. In order to make the distances consistent, we average all distances in an equivalence class (see Surface Variations) and store the average as the distance for every pair in the class. (This is only valid because the surface is symmetric.) Averaging has the effect of reducing variation in any particular distance due to random placement of the points. We found the distances before averaging to be

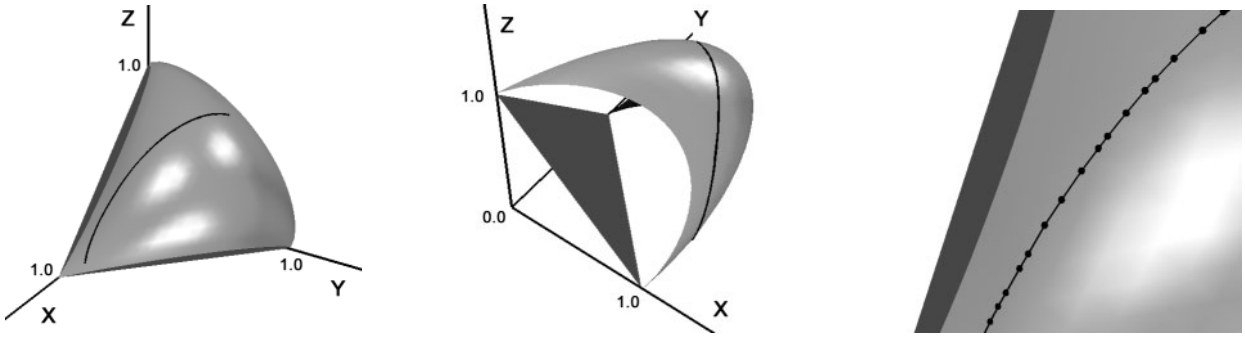


Fig. 5. (Left and center) Estimated shortest path between standard ($N = 30$) composition-by-count points $C_1 = (2, 1, 17)$ and $C_2 = (6, 10, 4)$ through the random graph. (Counts are in order Y, Z, X.) (Right) A close up of the path.

remarkably similar. Initial distance values range from 0 to approximately 2.1. The largest standard deviation for all equivalence classes is approximately 0.020 and the largest ratio of standard deviation to average distance in an equivalence class is 0.081. In other words, the standard deviation was, at worst, just over 8% of the average distance.

Scaling. We multiply each average distance by 10 and then round to whole numbers. This has the effect of spreading the distances. The final table values range from 0 to 21, with $d(C_A, C_T)$, the distance between conserved compositions C_A and C_T equal to 21.

APPLICATION TO COMPARISON OF TANDEM REPEATS

We used the new distance measure to analyze the relationships between tandem repeats found in the *C. elegans* genome. This approximately 100 megabase genome has been previously analyzed by the Tandem Repeats Finder (Benson, 1999) and contains approximately 25 000 tandem repeats. From this set, we selected 1175 repeats (1029 consensus patterns) with consensus pattern size ranging from 42 to 46 basepairs (bp). Many of these repeats are related as determined by alignment of the consensus sequences. Our hope was to refine/redefine their relationships by comparing the profiles.

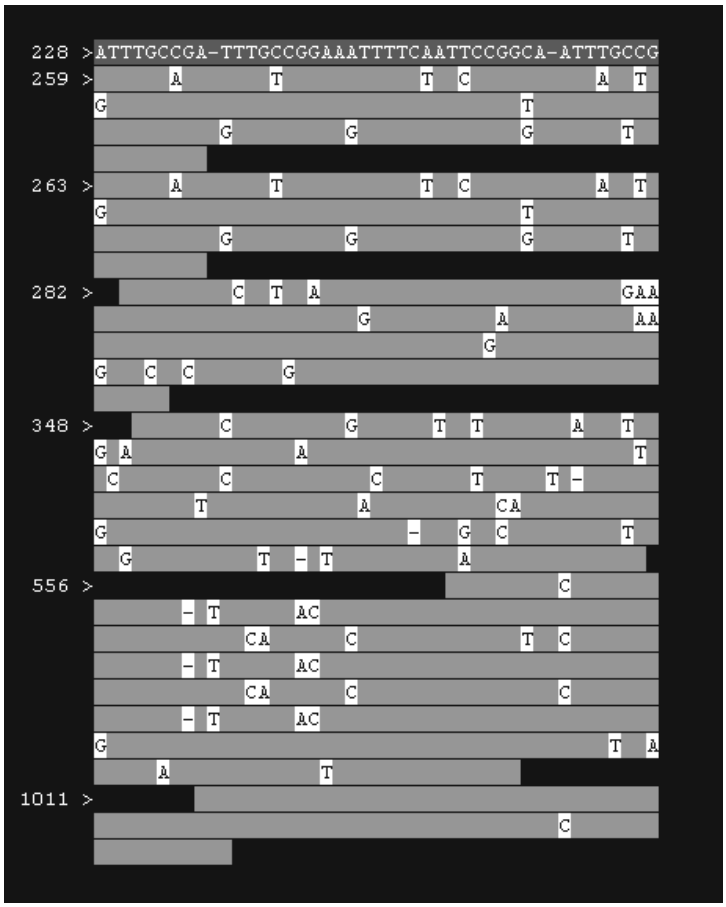
For each repeat, we use a multiple alignment of the individual copies. This is a 'star' alignment obtained by aligning each copy to the consensus, rather than the copies to each other (Figure 1 is an example). From the multiple alignment, we obtain the *standard profile*. Each column of the alignment is converted to its composition and the nearest standard composition ($k = 5, N = 10$, same as the distance table) by chord distance is determined. (This approximation allows us to use the precomputed distance table.) The standard profile consists, then, of a sequence of integers, one for each column of the multiple alignment, with each integer representing the standard composition closest to the true composition in that column.

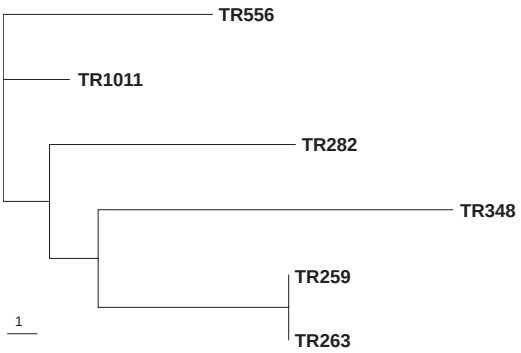
Comparison of profiles is performed using a cyclic alignment algorithm (Maes, 1990) which finds for profiles $S = s_1 s_2 \cdots s_n$ and $T = t_1 t_2 \cdots t_m$ the best scoring alignment of $S[i]$ versus T over all possible i for $i = 1 \ldots n$ where $S[i] = s_i s_{i+1} \ldots s_n s_1 \ldots s_{i-1}$. This is necessary because our initial profile (or consensus) position is defined by the first character of the repeat and it is frequently observed (see Figure 1) that related repeats do not start or end at the same relative positions. Because similar repeats may occur on opposite strands of DNA, for each pair of profiles, one is converted to its *reverse complement* and the pair is realigned. To obtain the reverse complement of a profile, we (1) reverse the order of the integer sequence and (2) replace integer i, which stands for composition $C_i = (f_A^i, f_C^i, f_G^i, f_T^i, f_-^i)$, by integer j, which stands for composition $C_j = (f_A^j, f_C^j, f_G^j, f_T^j, f_-^j)$, where C_j is obtained from C_i by exchange of the A/T frequencies and the C/G frequencies. Formally, $f_A^j = f_T^i$, $f_T^j = f_A^i$, $f_C^j = f_G^i$, $f_G^j = f_C^i$. The best score, S, obtained with either the normal or reverse complement alignment is the score for the profile pair.

Since we are using *distance* scoring for the alignments, identical profiles score zero and all scores are non-negative. We *normalize* the scores with the following formula:

$$S_{\text{norm}} = 200S/[(n+m)d(C_A, C_T)].$$

This accounts for size differences between the two profiles of length n and m and converts the score to a percentage where the distance $d(C_A, C_T)$ is included so that the normalized score equals 1 (1%) for two profiles of length 100 which differ only in the substitution of conserved compositions at one location. The normalization permits easy comparison between (1) alignments of repeat profiles





Profile distances

	*T R*259	*T R*263	*T R*282	*T R*348	*T R*556	*T R*1011
*T R*259	0.000	0.000	16.140	18.380	17.380	11.380
*T R*263	0.000	0.000	16.140	18.380	17.380	11.380
*T R*282	16.140	16.140	0.000	22.100	16.950	12.050
*T R*348	18.380	18.380	22.100	0.000	21.430	17.950
*T R*556	17.380	17.380	16.950	21.430	0.000	9.290
*T R*1011	11.380	11.380	12.050	17.950	9.290	0.000

Fig. 6. (Left) Alignment of six repeats with the same consensus sequence. (Right) Neighbor-joining tree based on profile distance for these repeats.

using the new distance table, and (2) alignments of consensus patterns using unit cost (number of differences) scoring.

Results

Consensus comparison can be misleading in several ways. It can over or under exaggerate the distance between repeats and it can give a distance of zero when the repeats are substantially different. Examples are presented in Figures 6 and 7. Figure 6, left, is a multiple alignment of five repeats all with the same consensus pattern (i.e. with consensus distance *zero*). The first two repeats (259, 263) are indeed identical, but the remaining three have many differences. (Repeat 556 shows internal repetition of mutations and may in fact be actively duplicating.) Profile comparison yields the much wider range of distances shown at the bottom of the figure. These normalized distances can be interpreted in the following way. A distance of 11 means that within the profile alignment, aligned compositions have, on average, a score which is 11% of the score for C_A versus C_T. A tree constructed by the Neighbor-Joining method (Saitou and Nei, 1987) using the profile distances is shown in Figure 6 right.

The tree in Figure 7, left, is also a Neighbor-Joining

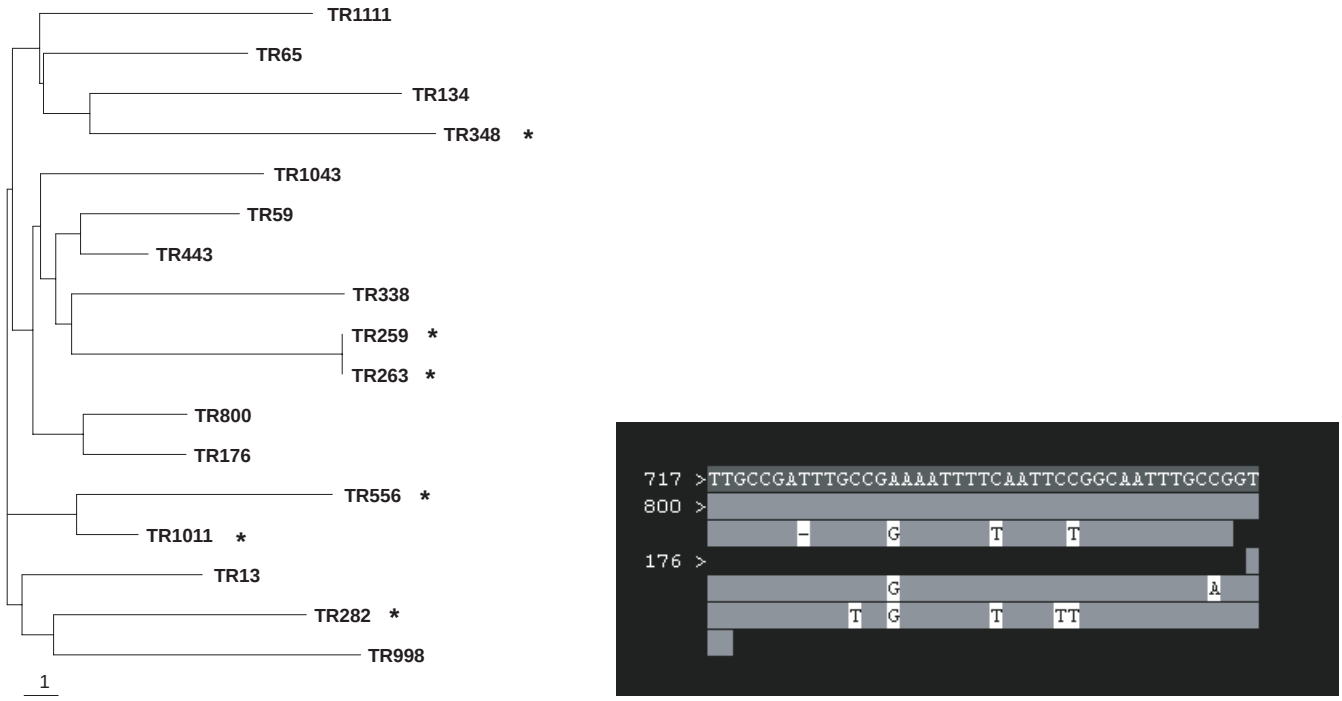


Fig. 7. (Left) Neighbor joining tree based on profile distance for a set of 17 repeats, including those in the previous figure (starred). Every pair of these repeats has a *consensus distance* of 4.65% (except pairs containing two starred sequences). (Right) Alignment of a close pair in the tree.

tree but constructed on the profile distances for a set of 17 repeats which include the six shown in Figure 6 (starred). Each pair of repeats in this tree (except pairs of starred repeats) has a *consensus distance* of 4.65%, thus there is no resolution of the relationships between these repeats using consensus distance. With profiles, we see that consensus comparison has both under and over exaggerated the distances between these repeats. Note that some of the new repeats are closer to each other and to the starred repeats than some of the starred repeats are to each other, even though the starred repeats all have a consensus distance of zero. An alignment of the closest non-starred pair (176, 800) showing obvious common mutations appears in Figure 7, right.

CONCLUSION

We have described a new distance measure for comparing sequence profiles. The measure is based on minimal path lengths along an entropy surface. Possible variations of the surface were described. Many other variations as well as other functions should be explored. We show how to approximate shortest paths using a random graph constructed on the entropy surface and Dijkstra's algorithm for all shortest distances to a single source. Profile comparison using the new measure was applied to tandem repeats from the *C. elegans* genome and two examples in this abstract show how the new measure is more accurate than comparing consensus sequences.

ACKNOWLEDGEMENTS

The auther would like to thank Alfredo Rodriguez for his invaluable help in producing the images for this paper and the tandem repeat analysis tools. The author would also like to thank Benjamin Kline and Craig Stevenson for their help in developing the random graph. It has been a pleasure working with all three. This work was supported in part by NSF grants CCR-0073081 and DBI-0090789.

REFERENCES

Ahlbrandt,C., Benson,G. and Casey,W. (2000) Minimal entropy probability paths. Manuscript.

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

Bennett,S., Lucassen,A., Gough,S., Powell,E., Undlien,D., Pritchard,L., Merriman,M., Kawaguchi,Y., Dronsfield,M., Pociot,F., Nerup,J., Bouzekri,N., Thomsen,A., Ronningen,K., Barnett,A., Bain,S. and Todd,J. (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene *Minisatellite locus*. *Nature Genet.*, **9**, 284–292.

Benson,G. (1997) Sequence alignment with tandem duplication. *J. Comput. Biol.*, **4**, 351–367.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Benson,G. (2000) Tandem cyclic alignment. In Amir,A. and Landau,G. (eds), *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Computer Science, 2089, Springer, pp. 118–130.

Cormen,T., Leiserson,C. and Rivest,R. (1990) *Introduction to Algorithms*. MIT Press.

Dijkstra,E.W. (1959) A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.

Edwards,A. and Cavalli-Sforza,C. (1964) Reconstruction of evolutionary trees. In Heywood,V. and McNeill,J. (eds), *Phenetic and Phylogenetic Classification*. Systematics Association, London.

Flèche,P.L., Hauck,Y., Onteniente,L., Prieur,A., Denoeud,F., Ramisse,V., Sylvestre,P., Benson,G., Ramisse,F. and Vergnaud,G. (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiology*, **1**, 2.

Fu,Y.-H., Pizzuti,A., Fenwick,Jr,R., King,J., Rajnarayan,S., Dunne,P., Dubel,J., Nasser,G., Ashizawa,T., DeJong,P., Wieringa,B., Korneluk,R., Perryman,M., Epstein,H. and Caskey,C. (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, **255**, 1256–1258.

Gribskov,M., Lüthy,R. and Eisenberg,D. (1990) Profile analysis. *Meth. Enzymol.*, **183**, 146–159.

Huntington's disease collaborative research group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.

Kececioglu,J. and Yu,J. (2001) Separating repeats in DNA sequence assembly. In *Proceedings of the 5th ACM Conference on Computational Molecular Biology (RECOMB 01)*.

Keim,P., Price,L., Klevytska,A., Smith,K., Schupp,J., Okinaka,R., Jackson,P. and Hugh-Jones,M. (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. *J. Bacteriol.*, **182**, 2928–2936.

Kullback,S. (1968) *Information Theory and Statistics*. Dover Publications.

Lalioti,M., Scott,H., Buresit,C., Rossier,C., Bottani,A., Morris,M., Malafosse,A. and Antonarakis,S. (1997) Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*, **386**, 847–851.

Lin,J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theor.*, **37**, 145–151.

Lyngso,R., Pederson,C. and Nielsen,H. (1999) Metrics and similarity measures for hidden Markov models. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology—ISMB99*. pp. 178–186.

Maes,M. (1990) On a cyclic string-to-string correction problem. *Information Processing Letters*, **35**, 73–78.

Nei,M., Tajima,F. and Tateno,Y. (1983) Accuracy of estimated phylogenetic trees from molecular data: II. gene frequency data. *J. Mol. Ecol.*, **19**, 153–170.

Owerbach,D. and Gabbay,K. (1993) Localization of a type 1 diabetes susceptibility locus to the variable tandem repeat region flanking the insulin gene. *Diabetes*, **42**, 1708–1714.

Phelan,C., Rebbeck,T., Weber,B., Deville,P., Ruttledge,M., Lynch,H., Lenoir,G., Stratton,M., Easton,D., Ponder,B., Albright,L., Larsson,C., Goldgar,D. and Narod,S. (1996) Ovarian cancer risk in BRCA1 carriers is modified by the HRAS1 variable number of tandem repeat (VNTR) locus. *Nature Genet.*, **12**, 309–311.

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

Smith,T. and Waterman,M. (1981) Comparison of biosequences. *Advances in Applied Mathematics*, **2**, 482–489.

Turk,G. (1990) Generating random points in triangles. In Glassner,A.S. (ed.), *Graphics Gems I*. Morgan Kaufmann.

Verkerk,A., Pieretti,M., Sutcliffe,J., Fu,Y., Kuhl,D., Pizzuti,A., Reiner,O., Richards,S., Victoria,M., Zhang,F., Eussen,B., van Ommen,G., Blonden,A., Riggins,G., Chastain,J., Kunst,C., Galjaard,H., Caskey,C., Nelson,D., Oostra,B. and Warren,S. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905–914.

Virtaneva,K., D'Amato,E., Miao,J., Koskiniemi,M., Norio,R., Avanzini,G., Franceschetti,S., Michelucci,R., Tassinari,C., Omer,S., Pennacchio,L., Myers,R., Dieguez-Lucena,J., Krahe,R., de la Chapelle,A. and Lehesjoki,A. (1997) Unstable minisatellite expansion causing recessively inherited myoclonus epilepsy, EPM1. *Nature Genet.*, **15**, 393–396.

Wong,A., Chan,S. and Chiu,D. (1993) A multiple sequence comparison method. *Bull. Math. Biol.*, **55**, 465–486.

Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.