1

Alignment-Free Comparison of TOPS Strings

David Gilbert, Francesc Rosselló, Gabriel Valiente and Mallika Veeramalai

ABSTRACT. TOPS diagrams are concise descriptions of the structural topology of proteins, and their comparison usually relies on a structural alignment of the corresponding vertex ordered and vertex and edge labelled graphs. Such an approach involves checking for the existence of subgraph isomorphisms, which is an NP complete problem even for this kind of graphs. Therefore, although there exist several algorithms for the alignment-based comparison of TOPS diagrams that are fast in practice, they have an exponential worst case complexity. Moreover, the alignment-based comparison of TOPS diagrams assumes conservation of contiguity between homologous TOPS diagram segments.

In this paper, we explore the alignment-free comparison of TOPS diagrams. We consider on the one hand similarity and dissimilarity measures based on subword composition of the sequences of secondary structure elements, thus neglecting contact map information, and on the other hand the Universal Similarity Metric from Kolmogorov complexity theory. Effectiveness of these alignment-free methods for TOPS diagrams comparison is assessed by cluster validation techniques.

1 Motivation

The number of known structures in the Protein Data Bank (PDB) [20] is increasing rapidly every year, with the PDB currently holding over 34,000 structures, as a result of efforts by the structural genomics consortium [3, 17] to populate protein fold space using high-throughput experimental technologies. This highlights the importance of the need for fast and reliable protein structure comparison methods, which can provide a better understanding of the structural and functional relationships between protein families.

There are, on the one hand, several methods that use detailed 3D structures for comparison, including SSAP [31, 37], STAMP [34], and DALI [15]. On the other hand, there are various methods that use more abstract topological descriptions of protein structure for comparison, like for instance

¡book title¿, 1–19.

VAST [27, 28], GRATH [13, 14], and TOPS [10, 41], as well as earlier approaches to find maximal common sequences of secondary structure elements in a pair of proteins [18]. Most of these methods model a protein structure as a sequence of secondary structure elements (SSEs), that is, of α-helices and β-strands, together with relationships like spatial neighborhood within the fold and approximate orientation, neglecting details of the structure like the lengths or the detailed structures of the SSEs themselves.

In this paper we focus on the comparison of TOPS diagrams, one of the most popular protein structure topological descriptions: see Section 2. We first recall the usual alignment-based comparison, which relies on the detection of least general common patterns [41]. Unfortunately, this method involves the detection of subgraph isomorphisms for vertex ordered and vertex and edge labelled graphs, which is an NP-complete problem. Moreover, the alignments preserve the order of the secondary structure elements in the protein structure, which means that this approach cannot detect similarity of structures when inter-domain motions occur.

Then we generalize to TOPS diagrams several alignment-free comparison methods that have been successfully used in the comparison of biological sequences and protein structures, and we assess and discuss their range of validity. A first group of methods rely on the comparison of subword frequency vectors of SSE sequences. The basic idea of this group of methods is that, the more similar two TOPS diagrams are, the greater is the similarity of their subword compositions. A second group of methods is based on data compression. The basic idea of this group of methods is that, the more similar two TOPS diagrams are, the more effective their joint compression is than their independent compression.

2 TOPS diagrams and TOPS patterns

TOPS diagrams [9, 12, 43, 44] provide a simple way to describe the *structural topology* of proteins, that is, their sequence of SSEs together with some information about the grouping of β-strands in β-sheets and about the orientation of SSEs.

In TOPS diagrams (for example the diagram for protein domain 2bopA0 in Figure 1), strands are represented by triangles and helices by circles, connected in a sequence from the amino (N) terminus to the carboxy (C) terminus. Secondary structure elements are considered to have a direction of 'up' (out of the plane of the diagram) or 'down' (into the plane of the diagram), implied in the way the connecting lines to the symbols are drawn: connections drawn to the edge of a symbol imply connection to the base and those drawn to the centre imply connection to the top, and the direction is that taken by the protein chain from N to C terminus. The direction

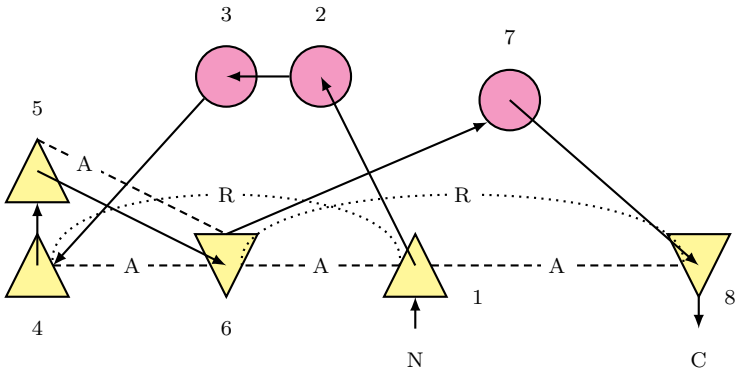

Figure 1. TOPS diagram for 2bopA0

information is duplicated for strands: upward pointing triangles have the direction 'up' and downward pointing ones the direction 'down'. The existence of hydrogen bond ladders between a pair of strands is indicated by a single H-bond in the TOPS representation, labelled as being parallel or anti-parallel, according to the relative directions of the two strands that it joins. In addition, TOPS diagrams also represent the chiralities of connections between two parallel strands within the same sheet and connections between long parallel helices. A more detailed description of TOPS diagrams can be found in [12].

A TOPS diagram can be more formally seen as a triple (S, H, C) where $S = S_1 \dots S_k$ is a sequence of length k of secondary structure elements, called a *TOPS string*, and H and C are relations over the SSEs, called respectively *H-bonds* and *chiralities*. In this description, an H-bond refers to a ladder of individual hydrogen bonds between adjacent strands in a sheet. We will refer to the *length* of a diagram as the length of the sequence S.

In our formalism, an SSE is represented by an H or an E, standing for helix and strand, respectively. But, since each SSE in a TOPS diagram is associated with a direction *up* or *down*, we must associate a direction with each one of these letters. This yields the alphabet $\{h, H, e, E\}$, where E stands for 'up' strand, e for 'down' strand, H for 'up' helix and h for 'down' helix.

Both H-bonds and chiralities are symmetric relations (non-directed arcs in the graph). H-bonds only occur between pairs of strands, and each H-

bond is associated with a relative direction $\delta \in \{P, A\}$, indicating whether the bond is between parallel (P) or anti-parallel (A) strands. Chiralities only occur between pairs of SSEs of the same type, and they are associated with handedness $\chi \in \{L, R\}$ (left and right respectively). We denote an H-bond relationship between two SSEs S_i and S_j in a TOPS string by (i, δ, j) and a chirality relationship by (i, χ, j).

The formal definition of a *TOPS diagram* is then a triple $D = (S, H_d, C_d)$ where, given $\Sigma = \{h, H, e, E\}$,

- $S = (S_1, \ldots, S_k)$, with $S_i \in \Sigma$, for every i;
- $H_d = \{(i, \delta, j) \mid S_i, S_j \in \{e, E\},\ \delta = P \text{ if } S_i = S_j,\ \delta = A \text{ if } S_i \neq S_j\}$;
- $C_d = \{(i, \chi, j) \mid S_i, S_j \in \{h, H\} \text{ or } S_i, S_j \in \{e, E\},\ \chi \in \{R, L\}\}$.

As an example, consider the TOPS diagram for 2bopA0 in Figure 1; we can 'stretch out' this diagram to give it a linear form, as shown in Figure 2, and represent it formally as 2bopA0 $= (S, H, C)$, where

- $S = (E, h, h, E, E, e, H, e)$;
- $H = \{(1, A, 6), (1, A, 8), (4, A, 6), (5, A, 6)\}$;
- $C = \{(1, R, 4), (6, R, 8)\}$.

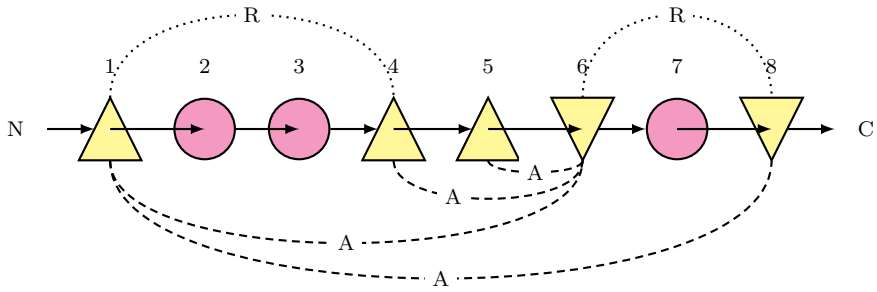

Figure 2. Linearised TOPS diagram for 2bopA0

We shall write such a TOPS diagram in a more compact form as follows:

```
2bopA0 NEhhEEeHeC 1:4R 1:6A 1:8A 4:6A 5:6A 6:8R
```

In this compact form, new letters are used to specify when an H-bond and a chirality occur between the same pair of SSEs: namely,

- A pair `i:j A` and `i:j L` is replaced by a single `i:j W`;
- A pair `i:j P` and `i:j L` is replaced by a single `i:j X`;
- A pair `i:j A` and `i:j R` is replaced by a single `i:j Y`;
- A pair `i:j P` and `i:j R` is replaced by a single `i:j Z`.

A TOPS pattern (or motif) is similar to a TOPS diagram, but is a generalisation which describes several diagrams conforming to some common topological characteristics. This generalisation is achieved by specifying the insertion of SSEs (and any associated H-bond and chiralities) into the sequence of secondary structure elements; indeed, a diagram is just a pattern where no inserts are permitted. The length of an insert is constrained to be within the range of the lengths of the sequences that can be inserted. So, the inserts are similar to wild cards with length constraints. We extend the definition of TOPS patterns given in [12] to permit such wild cards before the beginning and after the end of the sequence of SSEs.

More formally, a TOPS pattern is a triple (T, H, C) where T (referred to as a *T-pattern*) is a sequence

$$(n_0, m_0) - V_1 - (n_1, m_1) - V_2 - \cdots - (n_{k-1}, m_{k-1}) - V_k - (n_k, m_k)$$

comprising secondary structure elements indicated by V_i, and between each pair of consecutive SSEs an insert description, as well as insert descriptions (n_0, m_0) before V_1 and (n_k, m_k) after V_k. Each insert description is a pair (n, m) where n stands for the minimum and m for the maximum number of SSEs which can be inserted at that position. The range of n and m is from zero to the largest number N of SSEs in any TOPS diagram (currently, around 60). H is a set of H-bonds and C a set of chiralities, just as in the diagrams. The SSEs in a T-pattern are also associated with an 'up' or 'down' direction and represented by letters in $\{h, H, e, E\}$, as in TOPS diagrams.

So, the formal definition of a *TOPS pattern* is a structure $P = (T, H_p, C_p)$ where, given $\Sigma = \{h, H, e, E\}$,

- $T = (n_0, m_0) - V_1 - (n_1, m_1) - V_2 - \ldots - (n_{k-1}, m_{k-1}) - V_k - (n_k, m_k)$, with $V_i \in \Sigma$ and $n_i \leqslant m_i$, for every i;
- $H_p = \{(i, \delta, j) \mid V_i, V_j \in \{E, e\},\ \delta = P \text{ if } V_i = V_j,\ \delta = A \text{ if } V_i \neq V_j\}$;
- $C_p = \{(i, \chi, j) \mid V_i, V_j \in \{h, H\} \text{ or } V_i, V_j \in \{e, E\},\ \chi \in \{R, L\}\}$.

For example, a TOPS pattern that describes plaits, of which 2bopA0 is an instance, is given by Plait $= (T, H, C)$, where

- $T = (0, N) - E_1 - (0, N) - h_2 - (0, N) - E_3 - (0, N) - e_4 - (0, N) - H_5 - (0, N) - e_6 - (0, N)$;
- $H = \{(1, A, 4), (1, A, 6), (3, A, 4)\}$;
- $C = \{(1, R, 3), (1, R, 6)\}$.

We can write this in a more compact form as follows:

```
Plait N*E*h*E*e*H*e*C 1:3R 1:4A 1:6A 3:4A 4:6R
```

Figures 3 and 4 illustrate this in non-linear and linear form, respectively.

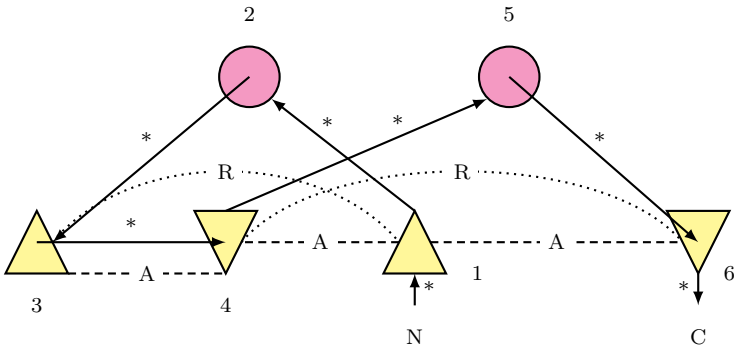

Figure 3. TOPS diagram for the plait motif

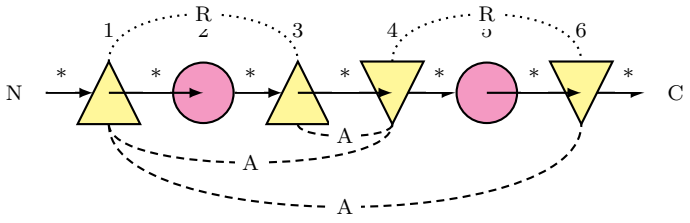

Figure 4. Linearised TOPS diagram for the plait motif

3 Alignment-based comparison

D. Gilbert and collaborators have designed a method to compare the similarity between two TOPS diagrams. This method is the basis of a topological level structure comparison system initially reported in [10] and since improved and updated as an on-line service [38], and linked to a sophisticated protein topology database [29].

Their method works by performing a structural alignment of the SSEs of the diagrams. In order to perform the alignment, they use a least general common pattern generated by a pattern discovery technique which they have designed. This in turn makes heavy use of their pattern matching method for TOPS diagrams: an efficient version is reported in [41], which is much more efficient than the initial algorithm described in [12]. The basis of the algorithm is based on repeatedly extending a pattern, and then attempting to match the pattern into all the examples in the input set. The procedure starts from the minimal, that is, empty, pattern.

Their algorithm discovers patterns of H-bonds and chiralities based on the properties of sheets for TOPS diagrams; it also derives the associated sequences of SSEs and insert sizes. Briefly, the algorithm attempts to discover a new sheet by finding, common to all the target set of diagrams—in the case of pairwise comparison just two diagrams, a (fresh) pair of strands, sharing an H-bond with a particular direction. Then it attempts to extend the sheet by repeatedly inserting a fresh strand which is H-bonded to one of the existing strands in the (current) sheet. The algorithm then finds all further H-bonds between all the members of the current sheet. The entire process is repeated until no more sheets can be discovered; any chirality arcs between the H-bonds in the pattern are then discovered by a similar process. The numbers of inserts between each strand in the pattern are then computed for all the patterns in the learning set, and the minimum and maximum size of the gaps in the corresponding insert positions in the pattern are thus found, and combined with the SSE sequence to give the T-pattern. The result is the least general common TOPS pattern characterising the target set of protein descriptions.

The distance measure M between two diagrams D_1 and D_2 is given by the normalised sum of the edit distances [22] of all the blocks plus a contribution from the extra (when compared with the pattern) H-bonds and chiralities in the diagrams. The distance between identical diagrams is zero; the larger the distance, the more dissimilar are the two diagrams.

Now, given TOPS diagrams $D_1 = (S_1, H_1, C_1)$ and $D_2 = (S_2, H_2, C_2)$, and a least general common pattern $Patt = (T, H, C)$ of them, we can make a structural alignment of D_1 and D_2 by matching T with S_1 and S_2. If $length(T) = n$, then there are $n + 1$ insert positions in the pattern,

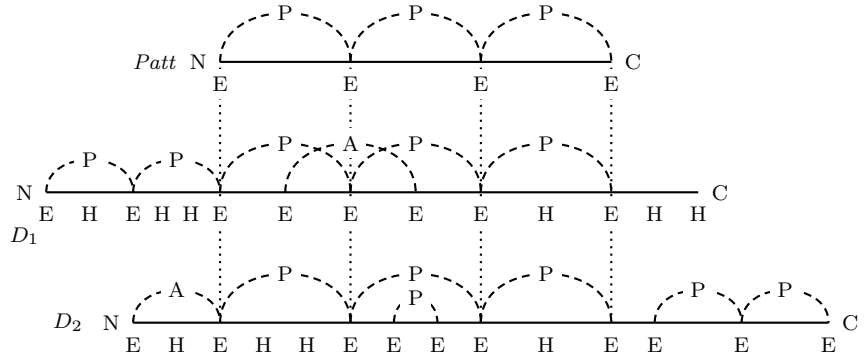

Figure 5. Making an alignment

corresponding to $n+1$ blocks of unaligned SSEs in S_1 and S_2. An example is illustrated in Figure 5.

Once this alignment has been produced, a distance measure between D_1 and D_2 can be defined by means of the normalised sum of the edit distances of the TOPS strings S_1 and S_2 with respect to this alignment, plus a contribution from the extra (when compared with the pattern) H-bonds and chiralities in the diagrams [10]. The distance between identical diagrams is zero; the larger the distance, the more dissimilar the two diagrams are.

More recently, another distance $d_{TOPS}(D_1, D_2)$ between D_1 and D_2 was proposed [11] (cf. also [8]) that simply measures how similar the least general common pattern is to the input TOPS diagrams. It is defined by

$$d_{TOPS}(D_1, D_2) = 1 - \frac{|Patt|}{\max(|D_1|, |D_2|)},$$

where $|Patt|$ is the *size* of the pattern as given by the total number of SSEs and arcs that it contains, $|D_1|$ is the size of diagram D_1, and $|D_2|$ is the size of diagram D_2. This measure varies from 0 (best) to 1 (worst).

EXAMPLE 1 *Consider the compact TOPS diagrams of domains 1qraA0, 4enl01, and 6xia00,*

```
1qraA0 NEheEhEhhEhEhC 1:4P 1:6P 3:4A 4:6R 6:9Z 9:11Z
4enl01 NEeehEhHhehEeHheHeHehHeHeHEHhC 1:26P 2:24P
       3:5A 5:9A 6:8L 9:12R 9:15P 11:12A 15:17Z
       17:19Z 19:22Z 22:24Z
```

```
6xia00 NhehHehHehHeEHheehHehHeHeHHHhhhhC 2:5Z 2:24P
       5:8Z 8:11Z 11:15Z 12:16A 15:19Z 19:22R 22:24Z
```

Their numbers in the SCOP classification of protein domains [30] are, respectively, c.37.1.8, c.1.11.1, and c.1.15.3. All three of them are alpha-beta proteins, but 4enl01 and 6xia00 are tim barrels.

A least general common TOPS pattern of each pair of these diagrams is given by the following table:

proteins	*pattern*
1hlb00 1jhgA0	`NhH*hH*hHC 3:5R`
1hlb00 5mbn00	`Nh*hhhhH*hHC 5:7R`
1jhgA0 5mbn00	`Nh*hH*hHC 3:5R`

This yields the distance values

$$d_{TOPS}(1hlb00, 1jhgA0) = 1 - \frac{7}{11} \approx 0.3637$$
$$d_{TOPS}(1hlb00, 5mbn00) = 1 - \frac{9}{11} \approx 0.1819$$
$$d_{TOPS}(1jhgA0, 5mbn00) = 1 - \frac{6}{10} \approx 0.4.$$

As it can be seen, the smallest value corresponds to the pair of goblins.

4 Alignment-free comparison

During the last twenty years, several alignment-free techniques for the comparison of strings have been developed [42]. A first group of such techniques is based on the comparison of word frequencies. Roughly speaking, every string x of length n over an alphabet A can be decomposed into $n - L + 1$ overlapping L-length words, and mapped to a vector

$$c_L^x = (c_{L,1}^x, c_{L,2}^x, \ldots, c_{L,K_L}^x) \in \mathbb{N}^{K_L}$$

of length $K_L = |A|^L$, where each $c_{L,i}^x$ is the number of occurrences in x of the ith (with respect to some fixed order on A^L) L-length word.

Since similar strings share word composition to some extent, the difference between two strings x and y can be quantified by means of some metric or some correlation index on the vectors c_L^x and c_L^y. Similarity and dissimilarity measures used for this purpose include the squared Euclidean distance

$$d_L^E(x, y) = \sum_{i=1}^{K_L} (c_{L,i}^x - c_{L,i}^y)^2,$$

the squared Mahalanobis distance

$$d_L^M(x,y) = (c_L^x - c_L^y)S^{-1}(c_L^x - c_L^y)^t$$

(where S stands for the covariance matrix of L-tuple occurrences), the angle between the vectors of frequencies,

$$d_L^{cos} = \arccos\left(\frac{\sum_{i=1}^{K_L} c_{L,i}^x c_{L,i}^y}{\sqrt{\sum_{i=1}^{K_L}(c_{L,i}^x)^2}\cdot\sqrt{\sum_{i=1}^{K_L}(c_{L,i}^y)^2}}\right),$$

and the linear correlation coefficient

$$LCC_L(x,y) = \frac{K_L\sum_{i=1}^{K_L} c_{L,i}^x c_{L,i}^y - \left(\sum_{i=1}^{K_L} c_{L,i}^x\right)\left(\sum_{i=1}^{K_L} c_{L,i}^y\right)}{\sqrt{K_L\sum_{i=1}^{K_L}(c_{L,i}^x)^2 - \left(\sum_{i=1}^{K_L} c_{L,i}^x\right)^2}\sqrt{K_L\sum_{i=1}^{K_L}(c_{L,i}^y)^2 - \left(\sum_{i=1}^{K_L} c_{L,i}^y\right)^2}}.$$

Notice that d_L^E, d_L^M and d_L^{cos} measure similarity, and therefore the largest the value they yield, the more different the strings' compositions are. On the other hand, LCC_L measures similarity, and hence it goes the other way round: it gives values between -1 (worst) and 1 (best).

To compare TOPS diagrams using this kind of techniques, we simply consider their TOPS strings and compare them as words over {H,h,E,e}, thus neglecting the information on H-bonds and chiralities. When computing vectors of subword occurrences, we shall consider the subwords ordered alphabetically with respect to

$$\mathtt{H} < \mathtt{h} < \mathtt{E} < \mathtt{e}.$$

For instance, we shall consider all length 2 words over this alphabet ordered as

HH, Hh, HE, He, hH, hh, hE, he, EH, Eh, EE, Ee, eH, eh, eE, ee.

EXAMPLE 2 *The protein domain 1qraA0 has TOPS string (after removing from its compact TOPS diagram the contact map and the initial N and the final C in the sequence of SSEs)*

EheEhEhhEhEh

This word has length $n = 12$ *and it can be decomposed into* $n - L + 1 = 11$ *overlapping words of length* $L = 2$*,*

`Eh he eE Eh hE Eh hh hE Eh hE Eh`

Therefore, the vector of numbers of length 2 word occurrences in this string is

$$c_2^{1\text{qraA0}} = (0, 0, 0, 0, 0, 1, 3, 1, 0, 5, 0, 0, 0, 0, 1, 0).$$

In a similar way, protein domains 4enl01 and 6xia00 have vectors of length 2 word occurrences

$$c_2^{4\text{enl01}} = (0, 3, 1, 4, 2, 0, 2, 2, 1, 1, 0, 2, 5, 3, 0, 1),$$

$$c_2^{6\text{xia00}} = (2, 2, 0, 6, 5, 3, 0, 2, 1, 0, 0, 0, 2, 5, 1, 1).$$

Their pairwise distances are as follows.

	(1qraA0, 4enl01)	*(1qraA0, 6xia00)*	*(4enl01, 6xia00)*
d_2^E	*90*	*139*	*51*
$d_2^{M,all}$	*49.8526*	*66.569*	*58.3037*
$d_2^{M,wr}$	*36.8002*	*52.009*	*45.1329*
d_2^{cos}	*1.3279*	*1.4783*	*0.7255*
LCC_2	*-0.276*	*-0.4439*	*0.4164*

In all cases, the most similar pair is the pair of tim barrels (4enl01, 6xia00), and the most different pair is (1qraA0, 6xia00).

The reader will notice that we have given two (squared) Mahalanobis distances in the last table, $d_2^{M,all}$ *and* $d_2^{M,wr}$*. They differ in the covariance matrix of numbers of length 2 words occurrences: the matrix used in* $d_2^{M,all}$ *has been computed from the set of vectors of length 2 words occurrence numbers of TOPS strings of all TOPS diagrams contained in the TOPS database, while* $d_2^{M,wr}$ *has been computed from the set of vectors of length 2 words occurrence numbers of all* pairwise different *TOPS strings of TOPS diagrams contained in the TOPS database.*

As another alternative to string comparison methods based on alignment, several metrics based on Kolmogorov complexity have been proposed recently in the literature [1, 2, 16, 23, 24, 40]. Roughly speaking, the *conditional Kolmogorov complexity* $K(x|y)$ of two strings x and y is the length of the shortest binary program P that computes x with input y [19]. Thus, $K(x|y)$ represents the minimal amount of information required to generate x by any effective computation when y is furnished as an input to the

computation. The *Kolmogorov complexity* $K(x)$ of a string x is defined as $K(x|\lambda)$, where λ stands for the empty string. Given a string x, let x^* denote the shortest binary program that produces x on an empty input; if there are more than one shortest program, x^* is the first one in alphabetic order. The *Kolmogorov complexity* $K(x, y)$ of a pair of objects x and y is the length of the shortest binary program that produces x and y and a way to tell them apart. More formal definitions of all these concepts and their main properties can be found in the textbook [26].

The most outstanding metric based on Kolmogorov complexity is the *Universal Similarity Metric* (actually, it only satisfies the axioms of metrics up to a certain additive precision) proposed in [24],

$$d(x, y) = \frac{\max\{K(x|y^*), K(y|x^*)\}}{\max\{K(x), K(y)\}},$$

which refines any other computable similarity metric, like for instance effective versions of Hamming distance, Euclidean distance, edit distances or alignment distances [24, Thm. VI.2]. This Universal Similarity Metric has been used successfully for instance to compute phylogenetic trees based on whole mitochondrial genomes [24, 6], to cluster SARS virus [6], to compare protein structures [21, 32], to reconstruct phylogenies from metabolic pathways [33], to classify languages [24], musical pieces [7, 6, 25], and images [36], to detect plagiarism in student assignments [4], and to cluster Russian literature [6].

Actually, since Kolmogorov complexities are non-computable in the Turing sense, the Universal Similarity Metric was not used in these applications as it stands, but approximations of it. The basis of these approximations is that $K(x)$ is intuitively the minimal amount of information required to generate x, that is, the shortest length of a compressed binary version of x, and therefore it is approximated by the length $C(x)$ of a compression of x. Furthermore, since $K(x, y) = K(xy)$ up to additive logarithmic precision [24], $K(x, y)$ can be approximated by the length $C(xy)$ of a compression of the concatenation of x and y. Finally, and since $K(x, y) = K(x) + K(y|x^*) = K(y) + K(x|y^*)$, again up to constant additive precision [26], the conditional complexity $K(x|y^*)$ can be approximated by $C(xy) - C(y)$, and $K(y|x^*)$ can be approximated by $C(yx) - C(x)$. This leads to the following approximation of the Universal Similarity Metric [21, 32], which is the one we use in this paper:

$$d(x, y) \approx \frac{\max\{C(xy) - C(x), C(yx) - C(y)\}}{\max\{C(x), C(y)\}}\,.$$

EXAMPLE 3 *Consider again the protein chains 6xia00, 4enl01, and 1qraA0, whose TOPS strings are, respectively,*

`NhehHehHehHeEHheehHehHeHeHHHhhhhC`

`NEeehEhHhehEeHheHeHehHeHeHEHhC`

`NEheEhEhhEhEhC`

Using the Perl interface `Compress::Zlib` *to the 'zlib' compression library to compress these strings, we obtain the following lengths (in bytes) of compressed TOPS strings,*

$$\begin{array}{ccc} C(\text{6xia00}) = 39 & C(\text{6xia00}\,\text{4enl01}) = 58 & C(\text{1qraA0}\,\text{6xia00}) = 51 \\ C(\text{4enl01}) = 43 & C(\text{4enl01}\,\text{6xia00}) = 58 & C(\text{4enl01}\,\text{1qraA0}) = 54 \\ C(\text{1qraA0}) = 31 & C(\text{6xia00}\,\text{1qraA0}) = 51 & C(\text{1qraA0}\,\text{4enl01}) = 53 \end{array}$$

from where

$$\begin{aligned} &d(\text{6xia00}, \text{4enl01}) \approx 0.4419, \\ &d(\text{6xia00}, \text{1qraA0}) \approx 0.5129, \\ &d(\text{4enl01}, \text{1qraA0}) \approx 0.5116. \end{aligned}$$

This is consistent with the fact that proteins 6xia00 and 4enl01 are both tim barrels, while 1qraA0 belongs to a different family, alpha-beta: the distance between 6xia00 and 4enl01 is lesser than the distance between any of them and 1qraA0.

When using the Universal Similarity Metric to compare protein chains, we have also considered representations where the contact map in the TOPS string is replaced by its adjacency matrix. So, for instance, we have also represented the protein chain 6xia00 by

```
NhehHehHehHeEHheehHehHeHeHHHhhhhC
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 Z 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 P 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 Z 0 0 0 0 0 Z 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 Z 0 0 0 0 0 Z 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 Z 0 0 0 0 0 0 Z 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 A 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 Z 0 0 0 0 0 0 0 Z 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 A 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 Z 0 0 0 0 0 0 R 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 R 0 0 0 0 Z 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 P 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 Z 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

and applied the compression algorithm to such a string.

5 Results

In order to assess the alignment-free comparison of TOPS strings, we have performed a series of experiments upon a subset of 36 protein domains drawn from PDB entries of five families (globin, alpha-beta, tim-barrel, all-alpha, all-beta), introduced in [5] and further studied in [21]. These proteins are:

globin 1eca00, 5mbn00, 1hlb00, 1hlm00, 1babA0, 1babB0, 1ithA0, 1mba00, 2hbg00, 2lhb00, 3sdhA0, 1ash00, 1flp00, 1myt00, 1lh200, 2vhbA0, 2vhb00.

alpha-beta 1aa900, 1gnp00, 6q21A0, 1ct9A1, 1qraA0, 5p2100.

tim barrel 6xia00, 2mnr01, 1chrA1, 4enl01.

beta 1cd800, 1ci5A0, 1qa9A0, 1cdb00, 1neu00, 1qfoA0, 1hnf01.

alpha 1cnpA0, 1jhgA0.

Protein chain 2vhb00 appears twice, as 2vhb00 and 2vhbA0, in order to test whether the two chains are detected by alignment-free comparison methods to be identical (and thus clustered together) or not. These proteins chains have TOPS strings ranging from 5 to 31 SSEs. The compact TOPS diagrams for the 36 protein chains are reproduced in Figure 6.

We have performed an all-against-all comparison using several alignment-free comparison methods and, in each case, we have clustered the resulting distance matrix using the Perl `Bio::Tree::DistanceFactory` implementation of the unweighted pair group method with arithmetic mean (UPGMA).

Since the gold standard for this dataset is available, we have chosen to assess the all-against-all comparison results using the F-measure [39], which combines precision and recall. The F-measure is defined as

$$F(C) = \sum_{t \in T} \frac{N_t}{N} \max_{C_k \in C} \frac{2P_{tk}R_{tk}}{(P_{tk} + R_{tk})},$$

where N_{tk} is the number of elements of class t within cluster C_k, precision is defined as $P_{tk} = N_{tk}/N_k$, and recall is defined as $R_{tk} = N_{tk}/N_t$.

As a reference, let us point out that the F-measure of the clustering of the Chew-Kedem dataset obtained from the alignment-based similarity measure d_{TOPS} defined in Section 3 is 0.9552.

We have performed an all-against-all comparison for words of length L, with $1 \leqslant L \leqslant 5$, using the squared Euclidean distance d_L^2, the Mahalanobis distances $d_L^{M,all}$ and $d_L^{M,wr}$ (see Example 2), the angle distance d_L^{cos}, and the linear correlation coefficient LCC_L. We have then clustered each resulting distance matrix. The values of F-measure we have obtained are given in the following table:

L	d_L^2	$d_L^{M,wr}$	$d_L^{M,all}$	d_L^{cos}	LCC_L
1	0.9216	0.9216	0.9216	0.8556	0.5982
2	0.9031	0.8577	0.8618	0.8687	0.5513
3	0.9031	0.8086	0.8254	0.8490	0.5235
4	0.8132	0.7682	0.7728	0.6708	0.5261
5	0.7830	0.6467	0.6520	0.6672	0.5363

As can be seen in Figure 7, the clustering of the Chew-Kedem dataset obtained using the squared Euclidean distance d_L^2 with $L = 2$, although not perfect, shows some biologically meaningful clusters. For instance, all beta proteins are clustered together, all but one alpha-beta proteins are also clustered together, and there is a cluster of all globbins together with two alpha proteins.

We have also performed an all-against-all comparison using the approximation of Kolmogorov complexity for two representations derived from TOPS diagrams, namely the TOPS string together with the contact map, and the adjacency graph of the TOPS diagram. The value of F-measure we have obtained in these cases are as follows.

representation	F-measure
string	0.5294
adjacency matrix	0.6094

6 Discussion

The alignment-free comparison of TOPS strings represents an interesting alternative to previous alignment-based methods. All-against-all comparison of TOPS strings using word frequencies yields clustering results that are close to those obtained using more detailed descriptions of protein structure. For instance, the clustering results obtained using words of length 2 and 3 have an F-measure of 0.9031, while the clustering results obtained in [21] by compressing a complete description of protein structure instead of a topological description only, have an F-measure of 0.9274, and the F-measure ranges between 0 and 1, the latter for a perfect clustering.

The fact that all-against-all comparison of TOPS strings using an approximation of Kolmogorov complexity yields worse clustering results than when using word frequencies, can be explained by TOPS strings being a compact representation of protein topological structure, which are thus hard to compress any further. As a matter of fact, the F-measure of 0.5294 obtained when compressing TOPS strings raises to 0.5768 when no compression at all is performed and the TOPS strings themselves is taken as an approximation of their Kolmogorov complexity.

Future work includes the refinement of the comparison of TOPS diagrams based on subword counts in two directions: on the one hand, the use of SVD to filter noise for large values of L, and on the other hand its extension to full TOPS strings, by taking into account H-bonds and chiralities possibly through the adjacency matrix they define. As far as the work on compression-based comparison of TOPS diagrams, it is necessary to search for more efficient compression algorithms that better approximate Kolmogorov complexities. Finally, we plan to extend the results of alignment-free comparison of TOPS strings to a larger dataset, such as the one proposed in [35].

Acknowledgements

The research described in this paper has been partially supported by Spanish DGES project BFM2003-00771 ALBIOM, by Spanish CICYT project TIN 2004-07925-C03-01 GRAMMARS, and by EU project INTAS IT 04-77-7178. MV was supported by a PhD studentship from the University of Glasgow.

BIBLIOGRAPHY

[1] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Phys. Rev. Lett.*, 88(4):048702, 2002.

[2] C. H. Bennett, P. Gács, M. Li, P. M. B. Vitányi, and W. Zurek. Information distance. *IEEE T. Inform. Theory*, 44(7):1407–1423, 1998.

[3] S. K. Burley, S. C. Almo, J. B. Bonanno, M. Capel, M. R. Chance, T. Gaasterland,

D. Lin, A. S. Caronali, F. W. Studier, and S. Swaminathan. Structural genomics: Beyond the Human Genome Project. *Nat. Genet.*, 23(2):151–157, 1999.

[4] X. Chen, B. Francia, M. Li, B. Mckinnon, and A. Seker. Shared information and program plagiarism detection. *IEEE T. Inform. Theory*, 50(7):1545–1551, 2004.

[5] L. P. Chew and K. Kedem. Finding the consensus shape for a protein family. *Algorithmica*, 38(1):115–129, 2003.

[6] R. Cilibrasi and P. M. P. Vitányi. Clustering by compression. *IEEE T. Inform. Theory*, 51(4):1523–1545, 2005.

[7] R. Cilibrasi, P. M. P. Vitányi, and R. de Wolf. Algorithmic clustering of music based on string compression. *Computer Music J.*, 28(4):49–67, 2004.

[8] M.-L. Fernández and G. Valiente. A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recogn. Lett.*, 22(6/7):753–758, 2001.

[9] T. P. Flores, D. M. Moss, and J. M. Thornton. An algorithm for automatically generating protein topology cartoons. *Protein Eng. Des. Sel.*, 7(1):31–37, 1994.

[10] D. Gilbert, D. Westhead, J. Viksna, and J. Thornton. A computer system to perform structure comparison using TOPS representations of protein structure. *J. Comput. Chem.*, 26(1):23–30, 2001.

[11] D. Gilbert, D. R. Westhead, and J. Viksna. Techniques for comparison, pattern matching and pattern discovery: From sequences to protein topology. In *Artificial Intelligence and Heuristic Methods in Bioinformatics*, pages 128–147. IOS Press, 2003.

[12] D. R. Gilbert, D. R. Westhead, N. Nagano, and J. M. Thornton. Motif-based searching in TOPS protein topology databases. *Bioinformatics*, 15(4):317–326, 1999.

[13] H. M. Grindley, P. J. Artymiuk, D. W. Rice, and P. Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.*, 229(3):707–721, 1993.

[14] A. Harrison, F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton, and C. Orengo. Recognizing the fold of a protein structure. *Bioinformatics*, 19(14):1748–1759, 2003.

[15] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233(1):123–138, 1993.

[16] E. Keogh, S. Lonardi, and C. A. Rtanamahatana. Toward parameter-free data mining. In *Proc. 10th ACM Int. Conf. Knowledge Discovery and Data Mining*, pages 206–215. ACM, 2004.

[17] S.-H. Kim. Shining a light on structural genomics. *Nat. Struct. Mol. Biol.*, 5:643–645, 1998.

[18] I. Koch, T. Lengauer, and E. Wanke. An algorithm for finding maximal common subtopologies in a set of protein structures. *J. Comput. Biol.*, 3(2):289–306, 1996.

[19] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Probl. Inform. Transm.*, 1(1):1–7, 1965.

[20] A. Kouranov, L. Xie, J. de la Cruz, L. Chen, J. Westbrook, P. E. Bourne, and H. M. Berman. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, 34(D):302–305, 2006.

[21] N. Krasnogor and D. A. Pelta. Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics*, 20(7):1015–1021, 2004.

[22] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[23] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 2001.

[24] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi. The similarity metric. *IEEE T. Inform. Theory*, 50(12):3250–3264, 2004.

[25] M. Li and M. Sleep. Melody classification using a similarity metric based on kolmogorov complexity. In *Proc. Conf. Sound and Music Computing SMC'04*, pages 126–129. IRCAM, 2004.

[26] M. Li and P. M. P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, 2nd ed. 1997.
[27] T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins*, 23(3):356–369, 1995.
[28] T. Madej and M. C. Mossing. Hamiltonians for protein tertiary structure prediction based on three-dimensional environment principles. *J. Mol. Biol.*, 233(3):480–487, 1993.
[29] I. Michalopoulos, G. M. Torrance, D. R. Gilbert, and D. R. Westhead. TOPS: An enhanced database of protein structural topology. *Nucleic Acids Res.*, 32:D251–D254, 2003.
[30] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247(4):536–540, 1995.
[31] C. A. Orengo and W. R. Taylor. A rapid method for protein structure alignment. *J. Theor. Biol*, 147:517–551, 1990.
[32] D. Pelta, J. R. Gonzales, and N. Krasnogor. Protein structure comparison through fuzzy contact maps and the universal similarity metric. In *Proc. 4th Conf. European Society for Fuzzy Logic and Technology and 11 Rencontres Francophones sur la Logique Floue et ses Applications (EUSFLAT-LFA, 2005)*, pages 1124–1129, 2005.
[33] F. Rosselló and G. Valiente. Alignment-free comparison of metabolic pathways. In *Proc. 10th Annual Int. Conf. Research in Computational Molecular Biology*, 2006. Poster abstract.
[34] R. B. Russell and G. J. Barton. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins*, 14(2):309–323, 1992.
[35] M. Sierk and W. Person. Sensitivity and selectivity in protein structure comparison. *Protein Sci.*, 13:773–785, 2004.
[36] N. Svangard and P. Nordin. Automated aesthetic selection of evolutionary art by distance based classification of genomes and phenomes using the universal similarity metric. In *Proc. Applications of Evolutionary Computing: EvoWorkshops 2004*, volume 3005 of *Lecture Notes in Computer Science*, pages 447–456. Springer-Verlag, 2004.
[37] W. R. Taylor and C. A. Orengo. Protein structure alignment. *J. Mol. Biol.*, 208(1):1–22, 1989.
[38] G. M. Torrance, D. R. Gilbert, I. Michalopoulos, and D. R. Westhead. Protein structure topological comparison, discovery and matching service. *Bioinformatics*, 21(10):2537–2538, 2005.
[39] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.
[40] J.-S. Varre, J.-P. Delahaye, and E. Rivals. The transformation distance: A dissimilarity measure based on movements of segments. *Bioinformatics*, 15(3):194–202, 1999.
[41] J. Viksna and D. Gilbert. Pattern matching and pattern discovery algorithms for protein topologies. In *Algorithms in BioInformatics*, volume 2149 of *Lecture Notes in Comput. Sci.*, pages 98–111, Berlin Heidelberg, 2001. Springer-Verlag.
[42] S. Vinga and J. Almeida. Alignment-free sequence comparison: A review. *Bioinformatics*, 19(4):513–523, 2003.
[43] D. R. Westhead, D. C. Hutton, and J. M. Thornton. An atlas of protein topology cartoons available on the world wide web. *Trends Biochem. Sci.*, 23(1):35–36, 1998.
[44] D. R. Westhead, T. Slidel, T. Flores, and J. M. Thornton. Protein structural topology: Automated analysis and diagrammatic representations. *Protein Sci.*, 8(4):897–904, 1999.

David Gilbert and Mallika Veeramalai
Bioinformatics Research Centre

Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, Scotland, UK
Email: drg@dcs.gla.ac.uk, mallika@brc.dcs.gla.ac.uk

Francesc Rosselló
Department of Mathematics and Computer Science
Research Institute of Health Science
University of the Balearic Islands
E-07122 Palma de Mallorca, Spain
Email: cesc.rossello@uib.es

Gabriel Valiente
Algorithms, Bioinformatics, Complexity and Formal Methods Research Group
Technical University of Catalonia
E-08034 Barcelona, Spain
Email: valiente@lsi.upc.edu

protein chain	TOPS string
1aa900	NEheEhEhEhEhC 1:4P 1:6P 3:4A 4:6R 6:8Z 8:10Z
1ash00	NhHhhhHhHC 5:7R 6:8R
1babA0	NhhHhHhHC
1babB0	NhHhhhhhHC
1cd800	NEeEeEeEheEEC 1:2A 2:7A 3:4A 3:9A 4:5A 6:7A 9:10A 9:11A
1cdb00	NEEeeEC 1:5P 2:3A 2:4A 4:5A
1chrA1	NEEhEhEhhEhEhEhEhHEeHC 1:19A 2:4Z 4:6Z 6:9Z 9:11Z 11:13Z 13:15Z 18:19A
1ci5A0	NeEeEeHEeEeC 1:10P 2:8A 3:4A 3:9A 4:5A 7:8A 9:10A
1cnpA0	NhHHhhHC
1ct9A1	NhhHeHeHeHhehhHHehhhHhhHHHhC 4:6Z 4:11P 6:8Z 8:11R 11:16Z
1eca00	NhHhhHhHhC 5:7R 6:8R
1flp00	NhHhHHhHhC 2:5L 5:7R 6:8R
1gnp00	NEheEhEhhEhEhC 1:4P 1:6P 3:4A 4:6R 6:9Z 9:11Z
1hlb00	NhHhhhhHHhHC 6:9R
1hlm00	NhHhHhHC 1:3R 3:5R 4:6R
1hnf01	NeEehEHeEeHEeC 1:12P 2:9A 3:5A 3:11A 5:7A 8:9A 11:12A
1ithA0	NhHhhhHHhHC 5:8R
1jhgA0	NhHhHhHC 1:3R 3:5R
1lh200	NhHhHhHhC 2:4L 4:6R
1mba00	NhHhhhHhHC 5:7R 6:8R
1myt00	NhhhHHHhHhC 6:8R
1neu00	NEEeeEeEehEheEC 1:4A 2:13P 3:10A 5:6A 5:12A 6:7A 8:10A 12:13A
1qa9A0	NeEeEeEeEeC 1:9P 2:7A 3:4A 3:8A 4:5A 6:7A 8:9A
1qfoA0	NEEeeEeHehEheEEC 1:4A 2:14P 3:10A 5:6A 5:12A 6:8R 8:10A 12:13A 12:14A
1qraA0	NEheEhEhhEhEhC 1:4P 1:6P 3:4A 4:6R 6:9Z 9:11Z
2hbg00	NhHhhHhHC 1:4R 4:6R 5:7R
2lhb00	NhHhhhhHhHC 6:8 R7:9R
2mnr01	NEhEhEhEhEhEhhEHEeHC 1:3Z 1:14P 1:17A 3:5Z 5:7Z 7:9Z 9:11Z 11:14R 14:16R 16:17A
2vhb00	NhHhHhHhC 5:7R
2vhbA0	NhHhHhHhC 5:7R
3sdhA0	NhhHhHHhHhC 3:6L 6:8R 7:9R
4enl01	NEeehEhHhehEeHheHeHehHeHeHEHhC 1:26P 2:24P 3:5A 5:9A 6:8L 9:12R 9:15P 11:12A 15:17Z 17:19Z 19:22Z 22:24Z
5mbn00	NhhhhhHhhHC 5:7R
5p2100	NEheEhEhhEhEhC 1:4P 1:6P 3:4A 4:6R 6:9Z 9:11Z
6q21A0	NEheEEhEhEhC 1:4P 1:5P 3:4A 5:7Z 7:9Z
6xia00	NhehHehHehHeEHheehHehHeHeHHHhhhhC 2:5Z 2:24P 5:8Z 8:11Z 11:15Z 12:16A 15:19Z 19:22R 22:24Z

Figure 6. TOPS strings for the 36 protein chains in the Chew-Kedem dataset.

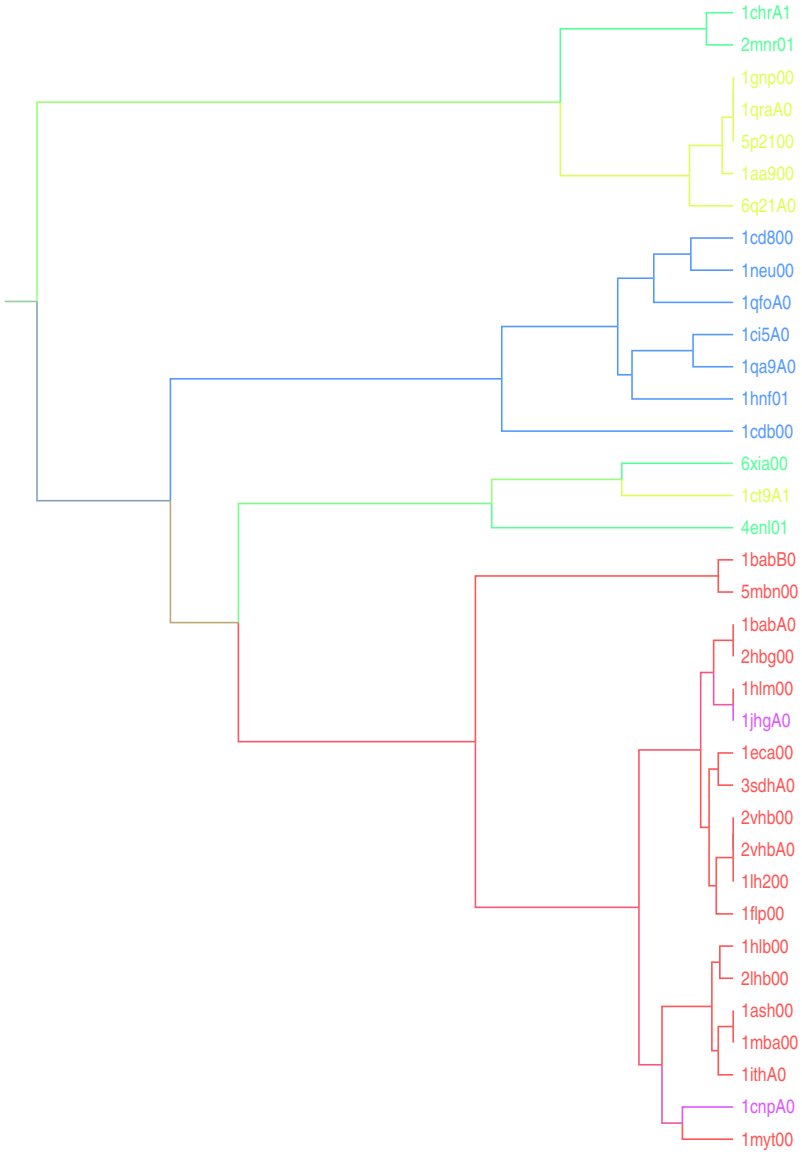

Figure 7. Clustering of the Chew-Kedem dataset based on the squared Euclidean distance d_L^2, for $L = 2$.