Biochimica et Biophysica Acta 1838 (2014) 1180–1190

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamem

A topological similarity measure for proteins ☆

Gabriell Máté [a], Andreas Hofmann [a], Nicolas Wenzel [a], Dieter W. Heermann [a,b,c,*]

[a] Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, Heidelberg, Germany
[b] The Jackson Laboratory, Bar Harbor, ME, USA
[c] Shanghai Institute of Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai, PR China

ARTICLE INFO

Article history:
Received 6 June 2013
Received in revised form 31 July 2013
Accepted 9 August 2013
Available online 10 September 2013

Keywords:
Protein similarity
Similarity measure
Protein flexibility
Protein structure
Persistent intervals
Jaccard index

ABSTRACT

We introduce a new measure for assessing similarity among chemical structures, based on well-established computational-topology algorithms. We argue that although the method considers geometry, it is more than a mere geometric similarity measure, as it takes into account, on different geometric scales, the important topological features of the compared structures. We prove that our measure is rigorous and complies with the proper mathematical requirements. We validate the method through comparing different configurations of simple zinc finger proteins and present an application on ligands binding to membrane-proteINS extracted from the Directory of Useful Decoys: Enhanced database and corresponding decoys. This article is part of a Special Issue entitled: Viral membrane proteins — Channels for cellular networking.

1. Introduction

Proteins, basic constituents of life, are probably one of the most important chemical structures for the living organisms. Their role ranges from replicating the DNA [1] through catalyzing numerous chemical reactions [2] to conferring stiffness for tissues [3]. Being responsible for a vast amount of biological processes, proteins play a crucial role in the formation and development of certain diseases, and as such, they are in the spotlight of drug design which often relies on investigating similarity relations among molecules [4].

There are two direct ways to assess similarity among chemical structures. One possibility is to determine a chemical similarity by converting the chemical formula into a graph, i.e., topology and compare topology as well as perhaps the chemical element similarity [5], but this may discard information regarding the folding of the investigated structures. The other possibility is to compare the geometry of the structures [6]. This approach in turn neglects important connectivity relations among the building elements of the chemical structure. In other words, both of these approaches are towards the extremes in the sense that one completely neglects the other.

While geometry is obviously an important factor, it is known that a large portion of proteins are relatively flexible structures and recently it has been understood that this property plays an important role in the binding process [7]. Therefore, flexibility is not a property which should be neglected when assessing the similarity of such molecules. In order to understand why generic comparison methods, relying exclusively on geometry or topology, may fail, first we need to understand the basic principles behind these methods.

In the present work we intend to establish a new approach for determining molecular similarity among chemical structures based exclusively on the physical configuration of these. We aimed to develop a method which takes into account the topological features and the geometry of the investigated structures. We achieved this by basing our method on well-established computational topology algorithms. We argue that although our method considers geometry, it is more than a method for calculating geometric similarity as it determines similarity based on observing prominent topological features on different geometric scales. Eventually, one could consider chemical or biological information as well.

We validate the method by calculating the similarities of different configurations of two zinc fingers connected by a flexible linker protein. Then we apply the method on the Directory of Useful Decoys: Enhanced (DUD-E) database [8], a docking database which contains many membrane-proteins, ligands binding to these and decoys specifically selected so that they do not bind.

This paper is organized as follows: First we shortly introduce the idea of geometric similarity, then we discuss the importance of topology and introduce our approach. After that we validate our method through comparing the zinc finger configurations. Last we present the application of the method on the DUD-E database.

☆ This article is part of a Special Issue entitled: Viral membrane proteins — Channels for cellular networking.

* Corresponding author at: Institute for Theoretical Physics, Heidelberg University, Philosophenweg 19, Heidelberg, Germany.
E-mail address: heermann@tphys.uni-heidelberg.de (D.W. Heermann).

0005-2736/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbamem.2013.08.019

2. Comparing proteins

The two generic approaches (comparing geometry and comparing the topology) may be the easiest way to determine similarity among molecules, however, when purely applying one or the other we discard important information. In order to demonstrate the flaws of these methods, we briefly introduce them.

2.1. Topological approach

Topology is the field of mathematics which investigates properties of objects which are invariant under certain deformations, i.e., stretching, bending — excluding breaking and tearing. Topological approaches (in fact all mathematical approaches) always require a good representation of the investigated objects. For instance, it would be really hard and thus unfeasible to represent a molecule with an abstract function.

Since, from the chemical point of view, the connectivity of the atoms is of crucial importance, these approaches intend to capture this aspect when representing a chemical structure. This information is easily stored by the chemical formula on the one hand but also by a more complex mathematical object called a *graph* [9]. Graphs are specially designed to capture connectivity information among different entities — atoms in the case of proteins, but they are suitable to represent any kind of structures composed of separable but interacting parts, commonly called *networks*. Usually, the interacting entities (e.g., representations of atoms) are referred to as vertices or nodes while connections between nodes (encoding chemical bonds, for instance) are represented by edges or links. Graphs can also be used to represent, for instance, computer networks [10], where the connected entities are the computers, on-line social networks [11], where nodes represent persons and edges represent friendships but also the complex connectivity characterizing the human brain [12].

In a purely topological, graph-theory based approach each atom of the investigated molecules is represented by a node and each bond is represented by an edge. The set of nodes and edges corresponding to a given molecule is a well-defined mathematical object, and there is a whole mathematical field built around these objects, called graph-theory.

Besides laying the foundations and defining the framework for handling graphs, graph-theory also provides the necessary measures and algorithms to compare graph-objects [13]. Without detailing these measures and methods, it is easy to understand now, that such an approach completely neglects any geometric or physical constraint since the representation of the data deals only with the connectivity information. Therefore, a purely topological approach could assess high similarity between a molecule and a physically and chemically incorrect copy of itself.

2.2. Geometric approach

Comparing molecules from a geometric point of view in turn supposes representing molecules as a form of volume. The easiest and perhaps the most realistic way to do this is by modeling each atom by a hard sphere with a radius corresponding to van der Waals radius of the atom. In this case, one can define geometric similarity as the Tanimoto or Jaccard measure of the volumes [14,6] calculated for the best alignment. This measure is defined as:

$$S_G(O_A, O_B) = \frac{V_A \cap V_B}{V_A \cup V_B}, \tag{1}$$

where O_A and O_B denote two different molecules, while V_A and V_B denote the volumes of O_A and O_B, respectively. The operation $V_A \cap V_B$ yields the section of the volumes while the operation $V_A \cup V_B$ yields the union of the volumes. Calculating the geometric similarity supposes that we previously calculated the best alignments, i.e., we tried to maximize this measure as a function of all possible rotations and translations. This is a computationally very costly procedure.

Although this measure performs very well when one is strictly interested in geometric similarity, proteins are flexible structures and flexibility turns out to be a very important property as it influences binding affinity [7] and function [15]. By calculating only geometric similarity we assess very reduced similarity between two different foldings of the same protein, which is obviously a bad result. On the other hand, geometric similarity is also sensitive to the difference in number of atoms.

3. Similarity and topological invariants

Based on the previous descriptions, it is clear that considering only topology or only geometry may lead to incorrect conclusions. There is a need for a method which is able to handle flexible structures and assess the correct similarity value even in complicated cases, for instance, when one compares two different distortions of the same object.

We build our similarity measure around two concepts: topology and physical constraints. Considering only topology would result in high similarity between a structure and its stretched version, which is an unwanted behavior. Note that considering physical constraints means that to some extent we are also interested in the geometry of the structures we want to compare.

A possible way to characterize topology is to record properties of the structures which are invariant under certain deformations of the object. Deformations which might fragment the structures (breaking, tearing, gluing, etc.) are excluded. In a more mathematical language, these deformations must correspond to continuous transformations of the topological space defined by the structures.

We will focus our attention to three quantifiable properties: the number of components which are independent from each other and connections only exist within components, the number of holes on the surfaces and the number of voids inside the structures. The field of algebraic topology has special names for these properties, they are called the Betti numbers of dimension zero, one and two, respectively, and they turn out to be very important topological invariants which help to distinguish between different topological spaces [16,17].

By comparing these quantities of two solid objects we can decide whether they have the same topology or not. But molecules are not solid objects. They are better described by the point-set defined by the coordinates of the atoms. Thus we need a method through which we can actually define what we mean by components, holes and voids.

To accomplish this, we, in fact, need to convert the point-set into a solid object. Therefore, imagine the following procedure: First, we take the point-set defined by the coordinates of the atoms and discard all the bond-information. From now on we will work only with these points. Next, we want to define a geometric relationship among the points. For this, we start growing spheres around each of them. Whenever two spheres mutually embed each-other's center we connect the centers of the spheres by a line/edge. Points connected by an edge are considered to belong to the same component. Any two points which are connected by a path through the existing edges are in the same component. As we increase the radii of the spheres we can record each event of connecting two previously disjoint components. By this we actually can follow how the number of components changes as a function of the radius. First each point is a separate component, while for a radius large enough each point is connected, and we end up with a single component.

The definition of holes and voids also stems from this process. In order to build a solid, beside points and lines, we need face and volume building blocks. For this, we will use the simplest polygon and polyhedron, namely the triangle and the tetrahedron. Whenever three edges form a triangle we consider not only the edges but also the face of the triangle. Similarly, whenever four triangles form a

tetrahedron, we consider the volume of the tetrahedron as solid volume. The described procedure is presented in Fig. 1 for a particular set of points.

Once the surfaces and volumes are defined we can proceed and count the holes and the voids. In fact, it is possible to register their number and also the number of components for every separate value of the radius of the spheres. This will be important in the next stage.

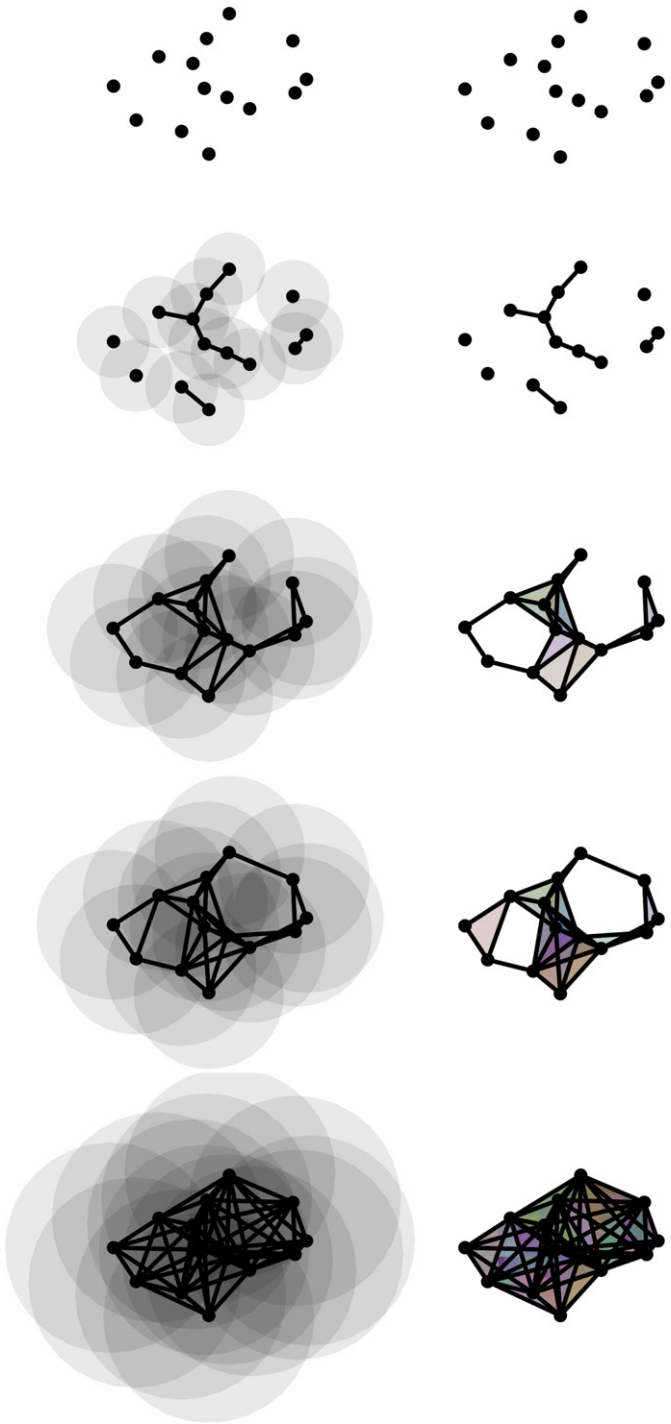

Fig. 1. Converting a point-set into a solid object. As the growing spheres mutually embed the center of each-other the corresponding centers are connected by an edge (as shown in the left column). Whenever a triangle/tetrahedron is formed, it is included in the solid as a face/volume element (illustrated in the right column).

3.1. A barcode representation of the structure

The results with respect to the change in the number of components, holes and voids throughout the previously described building process can be summarized in a single diagram in the following way:

- each instance of component, hole and void will be represented by a bar
- the position and length of a bar represents the "lifetime" of the corresponding component/hole/void
- the start point of the bar will correspond to the value of the radius at which the instance came into existence
- the end point of the bar will correspond to the value of the radius at which the instance ceased to exist.

The bars, in fact, are graphical representations of the intervals of the radii over which certain topological features (components, holes, voids) persist and they are called persistence intervals. The set of these bars characterizes how the topology of the object changes as we coarsen the representation of the structure and it can be viewed as a barcode of the topology on different scales. This representation was developed by Carlsson and his collaborators and a very good review of their work can be found in [18]. An example for such a barcode for a particular set of points can be seen in Fig. 2.

Note that for a given object we will have three different barcodes: one for components, one for holes and one for voids. In a mathematical terminology they are often referred to as dimension 0, dimension 1 and dimension 2 intervals, respectively.

To have a more physical understanding of the concept of components, holes and voids, imagine a regular rubber ball. The ball obviously has a single component and a void (usually filled with air) enclosed by the shell. If we poke a hole on the shell of the ball, we practically destroy what we in the context of this paper would call a void, as through the hole the air can escape. Now, in theory at least, we can grab the shell from the sides of the hole and stretch the rubber ball to a flat surface. In a mathematical language, we say that the ball with the hole is homeomorphic to a plane. Therefore, a single hole on a closed surface is in fact

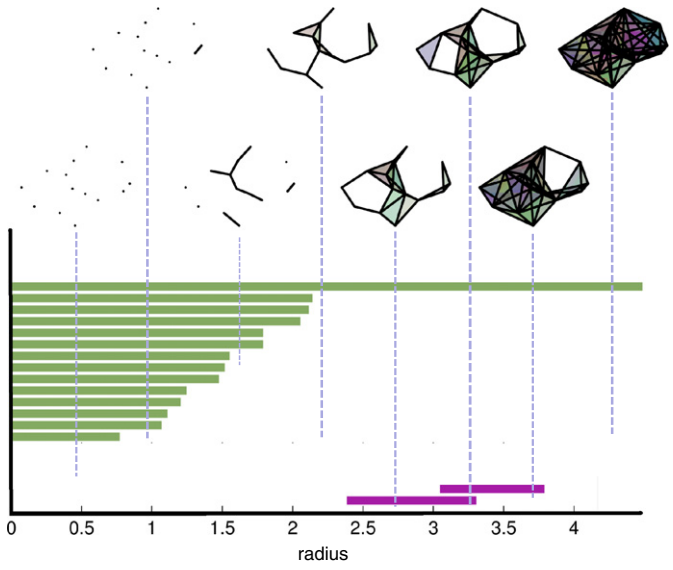

Fig. 2. Barcodes for a particular set of points in 2D. The horizontal axis represents the radius of the growing spheres. The green bars correspond to components while the purple ones correspond to holes. Persisting "features" are arranged on the vertical axis in an arbitrary order. On the top of the figure the procedure of connecting points is illustrated for a few values of the radius. Here, shaded faces signal formed triangles. Each triangle has a different color. Note how first each point constitutes a component, then as the radius increases the points start to connect to each other, thus the number of separate components decreases. Also note that the first hole forms at a radius value of around 2.4 while at a radius of 4.3 everything is connected, every hole is filled.

not a hole. Perforating the shell again and stretching from one of the holes results in an object homeomorphic with a plane with a "real" (topological) hole on it. Note that all the so far created objects had just a single component. In order to have two components we would need to cut the ball into two separated parts.

The bars/intervals for connected components (green lines in Fig. 2) are somewhat special as connected components unite as the radius increases. This process can be viewed as one of the connected components embeds the other one. Accordingly, the bar of the embedded component will end at the point where the component was embedded while the bar of the embedder component will continue until the latter will be embedded in another component. The role of embedded and embedder is arbitrary. It is easy to see that one of the bars for connected components will persist even at the highest values of the radius as there will always be at least one connected component, thus this bar can be neglected as it does not carry any information. For this reason, this bar may even be removed from the barcode.

Note that we are looking at the way the topology changes as we coarsen the representation of the structure we are investigating. By this we in fact implicitly consider geometric information without having to perform the expensive calculation of the best alignments. To understand how geometry is encoded in the barcodes let us return to the example with the ball. As already pointed out, this ball has a single connected component (its shell), no holes (otherwise the air would escape) and a single void inside the shell. Thus, there would be a single bar of a length corresponding to the diameter of the ball in the barcode representing voids. It is clear that if we change the geometry of the ball by flattening it for instance, we immediately would see the result of the change in geometry by the shrinkage of the bar representing the void inside the ball.

3.2. Similarity based on the barcodes

At this point we are able to calculate a barcode-representation of certain important topological features for a given structure. As we argued above, these barcodes also encode geometry. It is natural then to assess the similarity of two structures which may be of high complexity through comparing their barcodes, the latter being rather simple mathematical representation of the structures.

Since a barcode is in fact a set of bars, the first thing that comes to mind is the so-called Hausdorff distance [19] of the bar-sets. Although this approach would already provide an insight regarding similarity [7], the Hausdorff distance is a distance and not a similarity measure. It indicates the dissimilarity between two sets and its magnitude depends on the magnitude of the set-elements, that is, it is impossible to decide from the value of the Hausdorff distance of two sets whether the two sets are similar or not. We always have to provide a frame of reference. Although interpreting values of similarity measures defined on the interval [0,1] is not straightforward either, at least we know that values closer to one indicate high similarity, while values closer to zero mean reduced similarity.

Another classical way to compare sets is calculating their Jaccard or Tanimoto index (or measure) [14]. The Jaccard index is in fact the count of the elements present in both sets divided by the total number of elements, that is,

$$S_J(M,N)=\frac{|M\cap N|}{|M\cup N|}, \tag{2}$$

for any *nonempty* M and N sets. Unfortunately, in the case of the sets of the bars (the barcodes) it is not straightforward to apply the Jaccard similarity index since, for example, the coordinates are real valued numbers and bar-lengths may differ already because of experimental errors, thus deciding whether two bars from two different barcodes are equivalent or not is not a simple task. Also, we may consider two circles/rings similar even if their radius differs (different radius would mean different bar lengths). However, it is possible to define a measure based on the Jaccard index in the following way:

- we can calculate the Jaccard measure for every pair of intervals from two different barcodes
- for each bar from one barcode there exists a bar from the other barcode for which the Jaccard index is the highest
- we define our similarity measure as the average of these highest Jaccard measures.

Within a more mathematical framework, we can define this barcode-overlap similarity measure as

$$S_{BO}(A,B)=\frac{1}{|A|+|B|}\left[\sum_{a\in A}\sup_{b\in B}\frac{a\cap b}{a\cup b}+\sum_{b\in B}\sup_{a\in A}\frac{a\cap b}{a\cup b}\right], \tag{3}$$

where A and B denote two different barcodes while a and b denote different bars from barcodes A and B, respectively. Fig. 3 attempts to illustrate the calculation of this similarity. For the definition given in Eq. (3) it is possible to show that S_{BO} is a similarity measure in the mathematical sense (see proof in Appendix A).

Since we may encounter the case when there are no holes or voids in our structure, we need to extend the definition of our similarity measure so that we can handle these exceptions. This can be achieved by recognizing that an empty set is completely similar to another empty set. Therefore, we assign a value of 1 as the similarity between two empty barcodes. Also, note that the case when there are no bars in the barcode is quite different from the case when there are bars. Therefore, we assign a 0 similarity for this case. Compressing these in a mathematical formula, we get the following:

$$S_{BOE}(A,B)=\begin{cases} S_{BO}(A,B) & A\neq\emptyset \text{ and } B\neq\emptyset \\ 1 & A=\emptyset \text{ and } B=\emptyset \\ 0 & (A=\emptyset \text{ and } B\neq\emptyset) \text{ or } (A\neq\emptyset \text{ and } B=\emptyset). \end{cases} \tag{4}$$

Based on this definition, it is also possible to show that S_{BOE} is a proper similarity measure (see proof in Appendix B). The pseudocode describing the calculation of the S_{BOE} similarity measure is given in Algorithm 1.

The next question we are facing is how to unify the three similarity values we get from comparing the barcodes of connected components, holes and voids. Unfortunately, there is no unique way to do this. For example, we could take the average of the three numbers but we could also take the normalized Euclidean sum of the three, that is, summing the square of the three numbers, divide the outcome by three and then take the square root of the result. In fact, we could construct any method of unifying the values keeping in mind a single constraint: the method should not change the ordering of classification, that is, if a pair of objects is more similar in all the different barcodes then another pair of objects, the resultant unified similarity should be higher for the first pair. Mathematically speaking, we could apply any monotonically increasing function f which for any combination of input arguments from the range between zero and one would yield a result constrained to the same range, that is,

$$f:[0,1]^3\to[0,1]$$
$$f(x_1,x_2,x_3)\leq f(y_1,y_2,y_3),\ \forall x_i\leq y_i, x_j=y_j, i\neq j.$$

Important to note is that we must be consistent in our choice. It is not possible to compare two similarity values produced by two different forms of f. It makes even less sense to directly compare numerical values of geometric similarity to the values produced by S_{BOE} or any function of the latter.

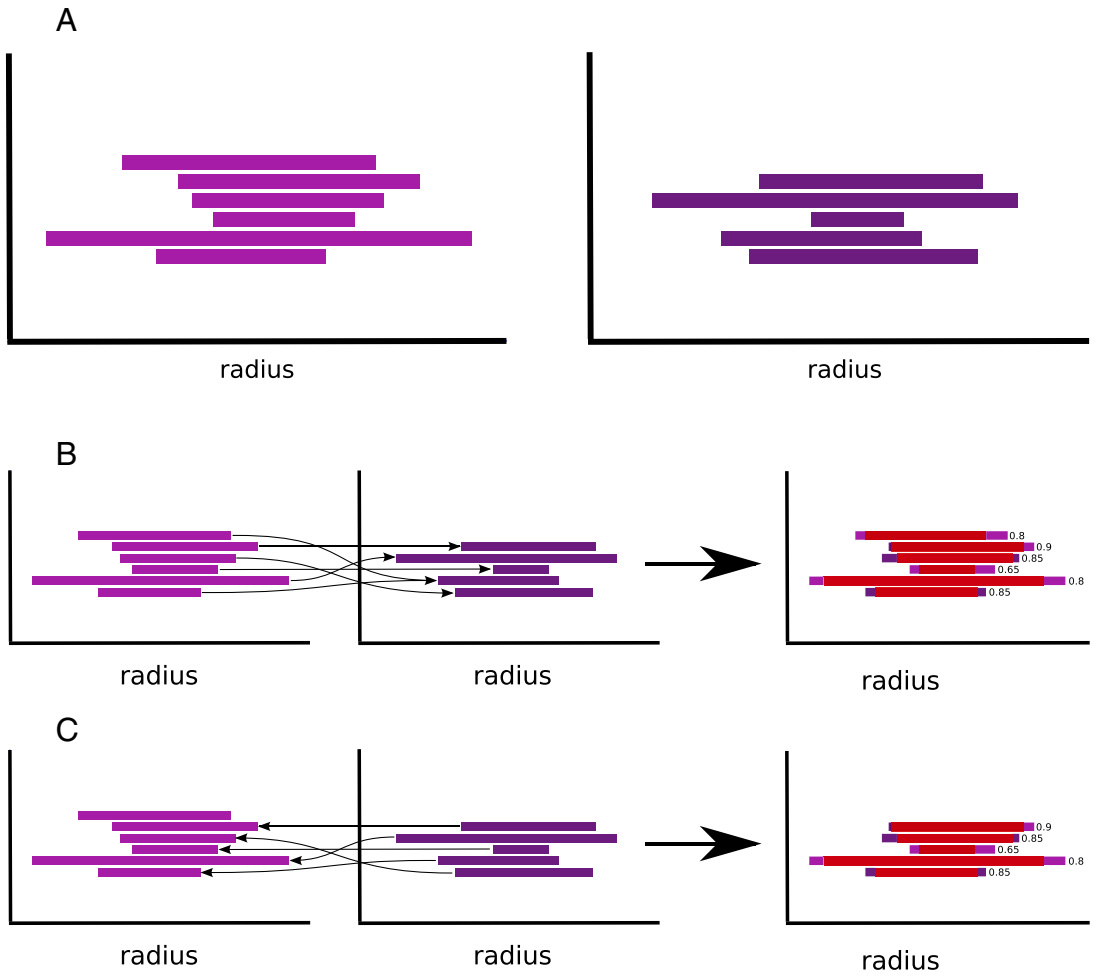

Fig. 3. In this figure we illustrate the calculation of the proposed similarity measure. Panel A presents two barcodes from two different molecules. Panel B illustrates the process of selecting for each bar from the first barcode from Panel A, those bars from the second barcode from Panel A for which the Jaccard index is the highest. Panel C illustrates this process for each bar from the second barcode. Overlaps are illustrated in red in the rightmost plots of Panels B and C. The (approximate) Jaccard indexes are also printed next to the illustrated overlaps. Our similarity measure is, in fact, the average of these indexes, which, in the presented case, would give a similarity of 0.8091.

For the sake of simplicity, we will define f as an average over the three arguments, that is:

$$f(x_1, x_2, x_3) = \frac{x_1 + x_2 + x_3}{3}, \tag{5}$$

and thus, we define the unified similarity measure as

$$S(O_A, O_B) = \frac{S_{BOE}(A_{cc}, B_{cc}) + S_{BOE}(A_{hl}, B_{hl}) + S_{BOE}(A_{vd}, B_{vd})}{3}, \tag{6}$$

where O_A and O_B denote two different objects/structures, A_{CC} and B_{cc} are the barcodes corresponding to connected components of the structures O_A and O_B, A_{hl} and B_{hl} are the barcodes for holes of the structures O_A and O_B, A_{vd} and B_{vd} are the barcodes representing voids of the structures O_A and O_B, respectively.

3.3. Validation of the method

As a validation of the method, here we calculate the S_{BOE} measures of the barcodes and the geometric similarity for four conformations of two zinc finger domains connected by flexible linker proteins, extracted from different configurations of CCCTC-binding factor (11-zinc finger protein) as presented in Fig. 4. Best overlaps among the configurations are illustrated in Fig. 5. We summarize the results of the comparison in Table 1.

The first observation in these comparisons is that configurations A and B have the smallest geometric resemblance. Comparing A against C yields a slightly larger geometric similarity, while this comparison gives a large value for the S_{BOE} similarity. Configurations B and C show

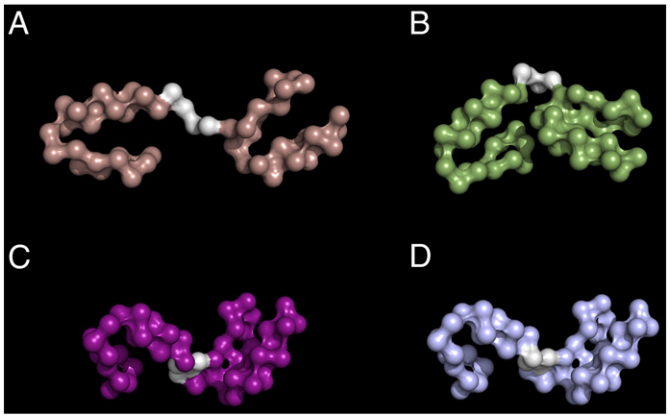

Fig. 4. Different configurations of two zinc finger proteins connected by a flexible linker protein. We use these configurations to validate our similarity measure (see Table 1) for the comparison.

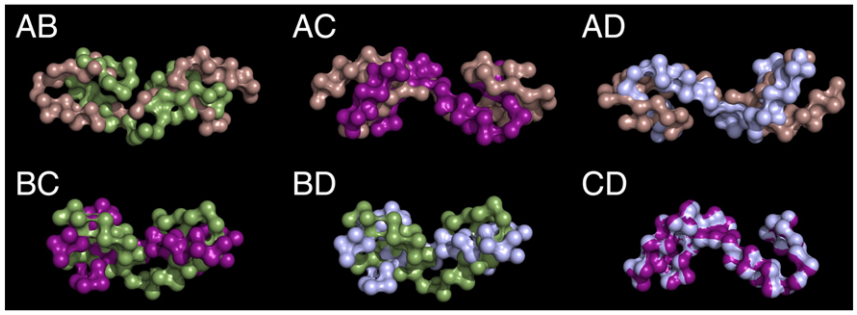

Fig. 5. Best alignments of the pairs of configurations from Fig. 4. For the values of the similarity measures for these pairs see Table 1.

a higher geometric similarity than A and C, while the S_{BOE} measure indicates a slightly reduced similarity compared to the A–C case.

Note that the C and D configurations are almost identical, they indeed have a very high geometric similarity, showing an increase of 0.3 compared to the A–C case, while the S_{BOE} similarity barely changes, ranking both pairs as very similar. Also note that comparing configuration B against any of the others consistently yields relatively reduced (but still high) S_{BOE} similarity, probably because of the particular features in the fold, while the geometric similarity of B to the other configurations is comparable to the values of similarity we get when comparing configuration A to the others, although A and B have the most reduced geometric similarity.

We remark that the S_{BOE} similarity of a value of around 0.6 configuration B shows when compared to the other configurations is considered relatively high as, comparing any of these configurations against a completely random configuration of comparable size returns a value averaging around 0.3 both for the S_{BOE} and the geometric similarity.

4. An Application

As a first application, we chose to compare the structures found in the Database of Useful (Docking) Decoys: Enhanced (DUD-E) database [8]. This database contains active ligands known to bind to given target-molecules and decoys which have geometries similar to those of the ligands, but they are chemically different. Decoys were selected from a vast amount of candidates and included in DUD-E based on two criteria. First, molecules were selected so that they have a high geometric similarity to one of the ligands, second, only those molecules were included in the database which were found to be inactive (molecules which do not bind to the target proteins — thus the name decoy).

We selected ligands grouped around fifteen target proteins (AA2AR, ABL1, ACE, ADA, ADRB1, AKT1, ALDR, ANDR, AOFB, BRAF, CAH2, COMT, CP2C9, DEF, HIVPR). Each of the ligands is known to bind at least to one of the targets. In this experiment we compare the ligands against the decoys from the same groups.

Although DUD-E was designed as a docking database, we use it for testing purposes. Since chemical differences must show up in the topology of the molecules, decoys and actives must present such differences. Therefore, it is a perfect sandbox for testing our similarity measure and to demonstrate that our measure picks up geometric similarity but it is not equivalent with it.

Table 1
Table presenting results for the geometric similarity and the introduced S_{BOE} similarity measure among the configurations of the zinc finger proteins presented in Fig. 4.

Tests	Geometric similarity	S_{BOE}
Config. A vs config. B	0.532	0.63853
Config. A vs config. C	0.641	0.97505
Config. A vs config. D	0.602	0.97791
Config. B vs config. C	0.669	0.63293
Config. B vs config. D	0.667	0.63615
Config. C vs config. D	0.955	0.98984

The calculations have two stages. First, there is a preprocessing step in which the barcodes are calculated. For this we used the Perseus software [20]. The calculated barcodes can be stored and there is no need to recalculate them at every comparison. A barcode, on average, can be calculated in roughly 12 s on a computer with a processor having a clock rate of 3.2 GHz. After barcodes for the present dataset of fifteen proteins were constructed, the similarities were calculated with a MATLAB script. Using the mentioned hardware, the runtime of the similarity calculations was 2684.2 s. Thus a comparison is performed in 0.0020629 s which roughly corresponds to 484 comparisons per second.

By looking at the distribution of the values of the geometric similarity (Fig. 6), we see that the values are centered around a well-defined mean value. It is possible to show, that these values actually follow a Gaussian distribution with a mean value of around 0.6.

Looking now at the distribution of the values of the S_{BOE} similarity illustrated in Fig. 7, we see that instead of having a single peak, a second peak may appear, which is caused by the unification of the different similarity values extracted from the barcodes of connected components, holes and voids as these different features may emphasize different aspects of the similarity. If we concentrate on the large peaks, we could say that the mean values are roughly around 0.75.

In Fig. 8, we present the values of the geometric similarity versus the values of the S_{BOE} index. Pairs for which the values are presented were selected so that the geometric similarity is among the largest values, roughly ranging from 0.8 to 0.95, well beyond the 0.55 average value. Note that almost all the corresponding S_{BOE} similarity values are also above their 0.75 average, most within the range between 0.75 and 0.92, that is, high geometric similarity implies high S_{BOE} values. Fig. 9, on the other hand, is prepared so that the values of the S_{BOE} similarity index are among the highest ones. Note that though the average of

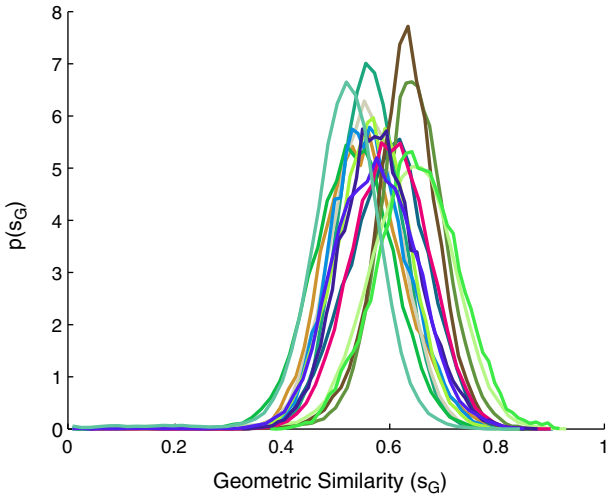

Fig. 6. Distribution of all the geometric similarity values among all the decoys and ligands from the 15 target proteins. Colors correspond to the different target proteins.

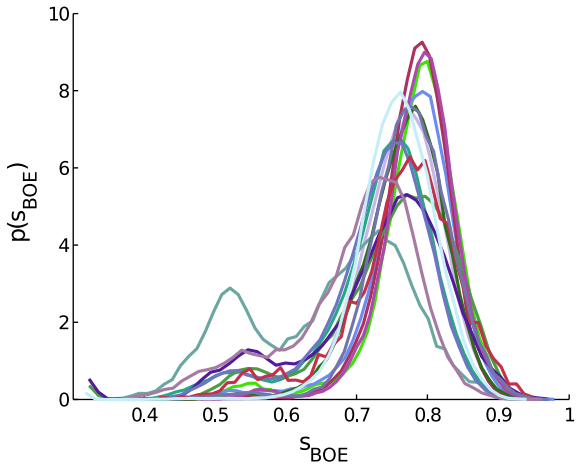

Fig. 7. Distribution of all the S_{BOE} values among all the decoys and ligands from the 15 target proteins. Colors correspond to the different target proteins.

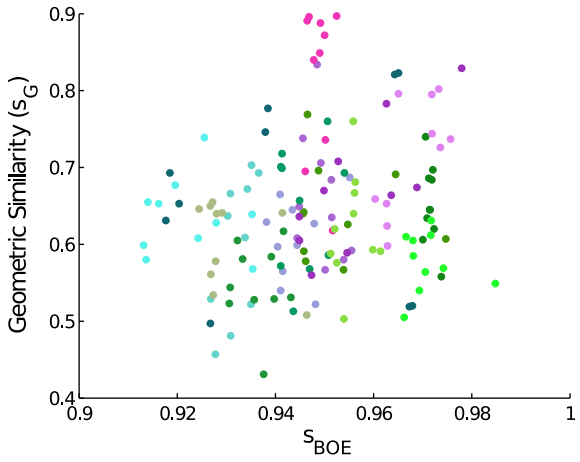

Fig. 9. Geometric similarity versus S_{BOE} for pairs of decoys and ligands. The pairs were selected so that their S_{BOE} similarity is among the largest values. Colors stand for the different target proteins.

the corresponding geometric similarity is higher than the global average, its values do not present such restriction as the values of the S_{BOE} similarity did in the previous case. This experiment clearly shows, that restricting geometric similarity to high values also restricts the S_{BOE} similarity index to higher values, while this is less true the other way around. This clearly indicates that the S_{BOE} measures more than the simple geometric similarity. In fact, it measures the similarity of the topological features on given geometric scales.

The same effect is also noticeable when looking at the ligands and the decoys themselves. In Fig. 10 we plotted pairs of ligands and decoys with the highest geometric similarity, while in Fig. 11 we show pairs of ligands and decoys for which both the geometric and S_{BOE} similarities rank high. As it can be seen, pairs geometrically resemble each other even when comparing them between the two figures. In Fig. 12, on the other hand, we show pairs with the highest S_{BOE} similarities. As it can be seen, these configurations are very different from the configurations seen in Figs. 10 and 11.

5. Discussion and conclusions

In this paper we introduced a novel similarity measure based on well-established computational topology algorithms. The measure was designed for assessing the similarity of different chemical structures but it may also be applicable in other fields. We proved that our definition is rigorous and it satisfies the mathematical requirements which are often neglected when new similarity measures are introduced.

Although the meaning of similarity is not clear-cut, being consistent in our choice is probably the most important principle to follow. It was easy to understand already based on our arguments that geometric similarity is not reliable and in certain cases it may fail. If we require consistency, mixing the values yielded by a given geometric similarity with other type of similarity measures is not viable. Therefore, we must construct similarity measures which, on the one hand, are proper measures, and, on the other hand, consider geometry, topology and other important factors at the same time. We believe that our method may be a good starting point for such an approach as we observed a logical path while welding geometry and topology and it is straightforwardly applicable when one is strictly interested in conformational similarities.

It is also important to form a good idea about the meaning of similarity. This is straightforward when it comes to geometry but it may not be so simple when one considers other features. As for our method, we would like to emphasize again, that our aim was to elaborate a measure which considers similarity beyond geometric resemblance, looks at the number of rings and other topological features, takes into account all the scales, but it is not scale invariant, while sticking to a rigorous mathematical background. Of course, the method is easily extendable. One of the first extensions one may want to implement is to input chemical information. This can be done, for instance, by introducing an extra “chemical-dimension” in the calculations.

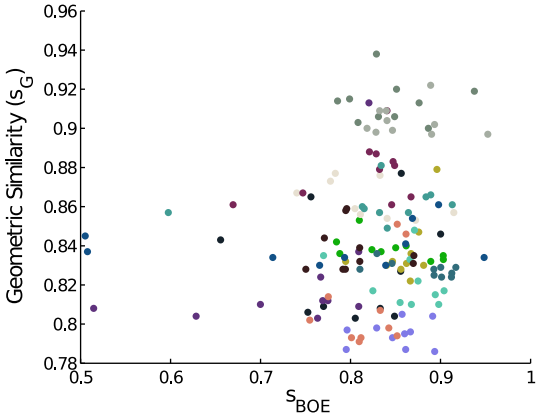

Fig. 8. Geometric similarity index versus S_{BOE} for pairs of decoys and ligands. The pairs were selected so that their geometric similarity is among the largest values. Colors stand for the different target proteins.

Acknowledgements

The authors would like to thank Lei Liu for the configurations of the two zinc finger molecules and for the useful informations he provided. They would also like to thank Yang Zhang for the interesting and useful discussions. Furthermore, GM gratefully acknowledges the support from the German Science Foundation (DFG) as member of the Research Training Group “Spatio/Temporal Probabilistic Graphical Models and Applications in Image Analysis”, grant GRK 1653, and the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences and the Institute for Theoretical Physics, all at the University of Heidelberg.

Appendix A

A.1. Definitions

Let A, B and C be three *nonempty* sets:

$$A = \{a | a = [a_s, a_e], a_s, a_e \in \mathbb{R}_+, a_s \leq a_e\} \tag{A.1}$$

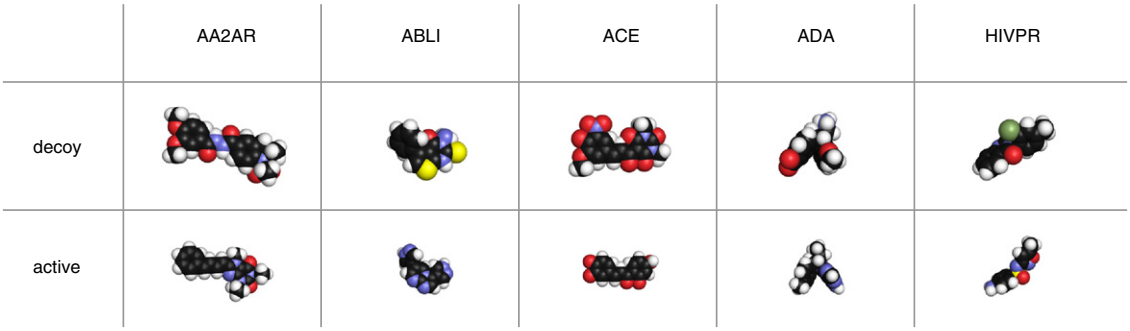

Fig. 10. Decoys and actives with the highest geometric similarity values.

$$B = \{b | b = [b_s, b_e], b_s, b_e \in \mathbb{R}_+, b_s \le b_e\} \tag{A.2}$$

$$C = \{b | c = [c_s, c_e], c_s, c_e \in \mathbb{R}_+, c_s \le c_e\}, \tag{A.3}$$

where $[x,y]$ denotes a closed interval with limits x and y.

Let $S_{BO}(A,B)$ be a mapping defined as:

$$S_{BO}(A,B) = \frac{1}{|A|+|B|}\left[\sum_{a\in A}\sup_{b\in B}\frac{|a\cap b|}{|a\cup b|} + \sum_{b\in B}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right]. \tag{A.4}$$

A.2. Aim

We intend to prove that S_{BO} is a proper similarity measure. According to [21] S_{BO} is a similarity relation if it satisfies the following conditions:

$$0 \le S_{BO}(A,B) \le 1 \tag{A.C5}$$

$$A = B \Rightarrow S_{BO}(A,B) = 1 \tag{A.C6}$$

$$S_{BO}(A,B) = S_{BO}(B,A) \tag{A.C7}$$

$$A \subseteq B \subseteq C \Rightarrow S_{BO}(A,C) \le S_{BO}(A,B) \tag{A.C8}$$

$$A \subseteq B \subseteq C \Rightarrow S_{BO}(A,C) \le S_{BO}(B,C). \tag{A.C9}$$

A.3. Proofs

Since for any $a \in A$ and $b \in B$ $|a \cap b|/|a \cup b|$ is between 0 and 1 for any A and B, $S_{BO}(A,B)$ will also be bounded by 0 and 1, thus Eq. (A.C5) is true.

For $A = B$ $\sup_{a\ inA}|a \cap b|/|a \cup b| = 1$ for any $b \in B$ and also $\sup_{b\ inB}|a \cap b|/|a \cup b| = 1$ for any $a \in A$. Therefore, $S_{BO}(A,B) = (|A| + |B|)/(|A| + |B| = 1)$, that is Eq. (A.C6) is true.

Condition (A.C7) is true by definition.

A.3.1. Condition (A.C8)

Proving $A \subseteq B \subseteq C \Rightarrow S_{BO}(A,C) \le S_{BO}(A,B)$.

Because of the relation $A \subseteq B \subseteq C$, the definition (A.4) for $S_{BO}(A,B)$ and $S_{BO}(A,C)$ can be rewritten in the following forms:

$$\begin{aligned} S_{BO}(A,B) &= \frac{1}{|A|+|B|}\left[\sum_{a\in A}\sup_{b\in B}\frac{|a\cap b|}{|a\cup b|} + \sum_{b\in B}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right] \\ &= \frac{1}{|A|+|B|}\left[|A| + \sum_{b\in A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|} + \sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right] \\ &= \frac{1}{|A|+|B|}\left[2|A| + \sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right], \end{aligned}$$

that is,

$$S_{BO}(A,B) = \frac{1}{|A|+|B|}\left[2|A| + \sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}\right], \tag{A.10}$$

similarly,

$$S_{BO}(A,C) = \frac{1}{|A|+|C|}\left[2|A| + \sum_{c\in C\setminus A}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]. \tag{A.11}$$

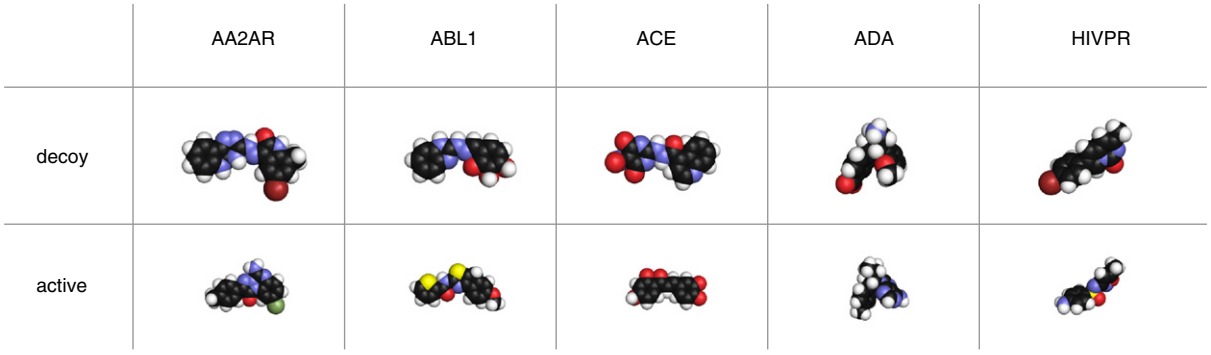

Fig. 11. Pairs of decoys and ligands with high geometric similarity and high S_{BOE} similarity values.

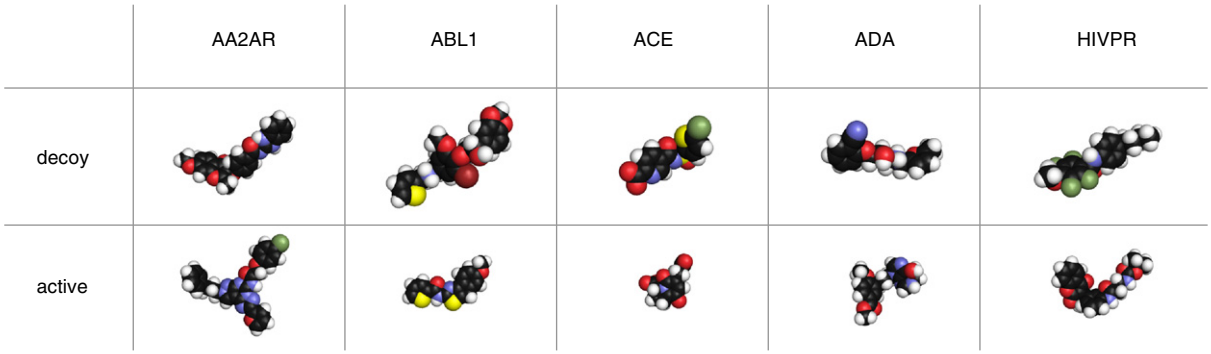

Fig. 12. Decoys and actives with the highest S_{BOE} similarity values.

Eq. (A.11) can be further rewritten:

$$S_{BO}(A,C)=\frac{1}{|A|+|C|}\left[2|A|+\sum_{c\in B\setminus A}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}+\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]. \tag{A.12}$$

Denoting

$$\sum_{b\in B\setminus A}\sup_{a\in A}\frac{|a\cap b|}{|a\cup b|}=:x, \tag{A.13}$$

we finally have

$$S_{BO}(A,B)=\frac{1}{|A|+|B|}[2|A|+x], \tag{A.14}$$

and

$$S_{BO}(A,C)=\frac{1}{|A|+|C|}\left[2|A|+x+\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]. \tag{A.15}$$

Then we can proceed as follows:

$$S_{BO}(A,C)\le S_{BO}(A,B)\Leftrightarrow \tag{A.16}$$

$$\frac{1}{|A|+|C|}\left[2|A|+x+\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]\le \tag{A.17}$$

$$\frac{1}{|A|+|B|}[2|A|+x]. \tag{A.18}$$

But since

$$\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\le|C|-|B|, \tag{A.19}$$

proving that

$$\frac{1}{|A|+|C|}(2|A|+x+|C|-|B|)\le\frac{1}{|A|+|B|}(2|A|+x) \tag{A.20}$$

is a stronger condition. From Eq. (A.20) we can proceed in the following way:

$$\frac{2|A|+x+|C|-|B|}{|A|+|C|}\le\frac{2|A|+|B|-|B|+x}{|A|+|B|}\Leftrightarrow \tag{A.21}$$

$$1+\frac{|A|+x-|B|}{|A|+|C|}\le1+\frac{|A|-|B|+x}{|A|+|B|}\Leftrightarrow \tag{A.22}$$

$$\frac{|A|-|B|+x}{|A|+|C|}\le\frac{|A|-|B|+x}{|A|+|B|}. \tag{A.23}$$

Inequality Eq. (A.23) is obviously true since $|A|+|C|\ge|A|+|B|$ as $A\subseteq B\subseteq C$. Thus Eq. (A.C8) is proved.

Appendix A.3.2. Condition (A.C9)

Here we prove that $A\subseteq B\subseteq C\Rightarrow S_{BO}(A,C)\le S_{BO}(B,C)$.

The formula for $S_{BO}(B,C)$ can be rewritten similarly to Eq. (A.11) form of $S_{BO}(A,C)$, that is,

$$S_{BO}(B,C)=\frac{1}{|B|+|C|}\left[2|B|+\sum_{c\in C\setminus B}\sup_{b\in B}\frac{|b\cap c|}{|b\cup c|}\right]. \tag{A.24}$$

Let

$$y:=\sum_{c\in C\setminus B}\sup_{b\in B}\frac{|b\cap c|}{|b\cup c|}. \tag{A.25}$$

Therefore, Eq. (A.24) simplifies to

$$S_{BO}(B,C)=\frac{1}{|B|+|C|}(2|B|+y). \tag{A.26}$$

Then, the statement we want to prove is

$$\frac{1}{|A|+|C|}\left[2|A|+x+\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\right]\le\frac{1}{|B|+|C|}(2|B|+y). \tag{A.27}$$

Note that since $A\subseteq B$ the following inequality holds:

$$\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\le\sum_{c\in C\setminus B}\sup_{a\in B}\frac{|a\cap c|}{|a\cup c|}, \tag{A.28}$$

that is,

$$\sum_{c\in C\setminus B}\sup_{a\in A}\frac{|a\cap c|}{|a\cup c|}\le y. \tag{A.29}$$

Therefore, if we can show that

$$\frac{1}{|A|+|C|}(2|A|+x+y)\le\frac{1}{|B|+|C|}(2|B|+y) \tag{A.30}$$

is true, then relation (A.27) will also hold.

From Eq. (A.13) we see that $x\le|B|-|A|$ and from Eq. (A.25) it results that $y\le|C|-|B|$. Since $|A|+|C|\le|B|+|C|$, one being the denominator on the left hand side of Eq. (A.30) the other being the denominator on the right hand side of the same equation, replacing y

on both sides of the equation with $|C| - |B|$, will have a larger contribution on the left hand side. Therefore, if the resulting inequality still holds, it means that Eq. (A.30) also holds and therefore Eq. (A.27) holds, too.

By carrying out the substitution we get the following:

$$\frac{2|A|+|C|-|B|+x}{|A|+|C|} \le \frac{2|B|+|C|-|B|}{|B|+|C|} \Leftrightarrow \tag{A.31}$$

$$\frac{|A|+|C|+|A|-|B|+x}{|A|+|C|} \le \frac{|B|+|C|}{|B|+|C|} \Leftrightarrow \tag{A.32}$$

$$1+\frac{|A|-|B|+x}{|A|+|C|} \le 1 \Leftrightarrow \tag{A.33}$$

$$\frac{|A|-|B|+x}{|A|+|C|} \le 0. \tag{A.34}$$

Since $|A| + |C| > 0$, Eq. (A.34) is equivalent with $|A| - |B| + x \le 0$. But from Eq. (A.13) we already saw that $x \le |B| - |A|$, therefore, our last statement is true which means that Eq. (A.27) is true, that is, Eq. (A.C9) is true.

By this we proved that S is a proper similarity measure.

Appendix B

B.1. Definitions

As the Jaccard index is not defined for empty sets, here we extend the proof presented in Appendix A to the case which allows comparing empty sets. Since the empty set is similar to itself, we define the similarity of two empty sets as total similarity, taking the value of 1. Furthermore, since the empty set is totally different from any non-empty set, we assign the value of 0 to the similarity between the empty set and any nonempty set. In mathematical terms, this means that we need to prove that the measure defined as

$$S_{BOE}(A,B) = \begin{cases} S_{BO}(A,B) & A\neq\varnothing \text{ and } B\neq\varnothing \\ 1 & A=\varnothing \text{ and } B=\varnothing \\ 0 & (A=\varnothing \text{ and } B\neq\varnothing) \text{ or } (A\neq\varnothing \text{ and } B=\varnothing) \end{cases} \tag{B.1}$$

is a similarity measure.

B.2. Proof

The proofs for the conditions (A.C5), (A.C6) and (A.C7) are relatively simple:

- Since $S_{BO} \in [0,1]$, S_{BOE} is also constrained to the interval [0,1], therefore, Eq. (A.C5) is true.
- If $A = B$, this means that both are either empty or not. If both are empty, then according to Eq. (B.1) definition $S_{BOE}(\varnothing,\varnothing) = 1$. If they are not empty then $S_{BOE}(A,B) = S_{BO}(A,B)$. But we already saw that if $A = B$ then $S_{BO}(A,B) = 1$. Therefore, Eq. (A.C6) is true.
- S_{BOE} is symmetric by definition, that is Eq. (A.C7) is true.

B.2.1. Proving Eqs. (A.C8) and (A.C9)

In order to show that Eqs. (A.C8) and (A.C9) both hold, we need to consider four different cases of the condition $A \subseteq B \subseteq C$:

$$A\neq\varnothing,\ B\neq\varnothing,\ C\neq\varnothing \tag{B.C2}$$

$$A=\varnothing,\ B\neq\varnothing,\ C\neq\varnothing \tag{B.C3}$$

$$A=\varnothing,\ B=\varnothing,\ C\neq\varnothing \tag{B.C4}$$

$$A=\varnothing,\ B=\varnothing,\ C=\varnothing. \tag{B.C5}$$

We now go through these different cases.

- in case Eq. (B.C2) is obviously the case when $S_{BOE} \equiv S_{BO}$, therefore, both Eqs. (A.C8) and (A.C9) hold in this case.
- in case Eq. (B.C3) $S_{BOE}(A,B) = 0$, $S_{BOE}(A,C) = 0$, $S_{BOE}(B,C) = S_{BO}(B,C) \in [0,1]$. Therefore, condition (A.C8) is equivalent with $0 \le 0$, while condition (A.C9) can be written as $0 \le S_{BO}(B,C)$. It is evident that both of these affirmations hold, therefore, both conditions are satisfied.
- in case Eq. (B.C5) $S_{BOE}(A,B) = 1, S_{BOE}(A,C) = 0, S_{BOE}(B,C) = 0$. Therefore, condition (A.C8) is equivalent with $0 \le 1$, while condition (A.C9) can be written as $0 \le 0$. These affirmations again hold, therefore, both conditions are satisfied.
- in case Eq. (B.C5) $S_{BOE}(A,B) = 1, S_{BOE}(A,C) = 1, S_{BOE}(B,C) = 1$. Therefore, condition (A.C8) is equivalent with $1 \le 1$, while condition (A.C9) can be written as $1 \le 1$. Since these are all true, the original conditions are again satisfied.

Based on the previous points, we see that if S_{BO} is a proper similarity, then S_{BOE} is also a similarity measure.

Algorithm 1. Calculating the S_{BOE} similarity

Algorithm 1 Calculating the S_{BOE} similarity

```
procedure S_BOE(A, B)
    if A = ∅ AND B = ∅ then
        return 1
    else if (A = ∅ AND B ≠ ∅) OR (A ≠ ∅ AND B = ∅) then
        return 0
    else
        pos ← 1
        for a ∈ A do                ▷ calculating the first sum from equation (3)
            jac[pos] ← 0
            for b ∈ B do
                if Jaccard(a, b) > jac[pos] then
                    jac[pos] ← Jaccard(a, b)
                end if
            end for
            pos ← pos + 1
        end for
        for b ∈ B do                ▷ calculating the second sum from equation (3)
            jac[pos] ← 0
            for a ∈ A do
                if Jaccard(a, b) > jac[pos] then
                    jac[pos] ← Jaccard(a, b)
                end if
            end for
            pos ← pos + 1
        end for
        sim ← 0
        for i ← 1, pos − 1 do                ▷ averaging the results
            sim ← sim + jac[i]
        end for
        return sim/(pos − 1)
    end if
end procedure
procedure Jaccard(a, b)     ▷ Calculates the Jaccard index of two bars
    s ← a ∩ b
    u ← a ∪ b
    return |s|/|u|
end procedure
```

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular biology of the cell, Garland DNA Replication, Repair, and Recombination, 4th edition, 2002. (Ch. 5, URL http://www.worldcat.org/isbn/0815332181).

[2] A. Kohen, J.P. Klinman, Enzyme catalysis: beyond classical paradigms, Acc. Chem. Res. 31 (7) (1998) 397–404, http://dx.doi.org/10.1021/ar9701225(URL http://pubs.acs.org/doi/abs/10.1021/ar9701225).

[3] P. Bornstein, J.F. Ash, Cell surface-associated structural proteins in connective tissue cells, Proc. Natl. Acad. Sci. 74 (6) (1977) 2480–2484(URL http://www.pnas.org/content/74/6/2480.abstract).

[4] H. Lin, M.F. Sassano, B.L. Roth, B.K. Shoichet, A pharmacological organization of g protein-coupled receptors, Nat. Methods 10 (2) (2013) 140–146, http://dx.doi.org/10.1038/nmeth.2324 (URL http://dx.doi.org/10.1038/nmeth.2324).

[5] S. Vishveshwara, K.V. Brinda, N. Kannan, Protein structure: insights from graph theory, J. Theor. Comput. Chem. 1 (1) (2002) 187–212 (URL http://mbu.iisc.ernet.in/~vishgp/pdf/graph_review_JTCC.pdf).

[6] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (6) (1998) 983–996, http://dx.doi.org/10.1021/ci9800211 (URL http://dx.doi.org/10.1021/ci9800211).

[7] C.J. Feinauer, A. Hofmann, S. Goldt, L. Liu, G. Máté, D.W. Heermann, Chapter three-zinc finger proteins and the 3d organization of chromosomes, in: R. Donev (Ed.), Organisation of Chromosomes, Advances in Protein Chemistry and Structural Biology, vol. 90, Academic Press, 2013, pp. 67–117, http://dx.doi.org/10.1016/B978-0-12-410523-2.00003-1, (URL http://www.sciencedirect.com/science/article/pii/B9780124105232000031).

[8] M.M. Mysinger, M. Carchia, J.J. Irwin, B.K. Shoichet, Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, J. Med. Chem. 55 (14) (2012) 6582–6594, http://dx.doi.org/10.1021/jm300687e(URL http://pubs.acs.org/doi/abs/10.1021/jm300687e).

[9] K. Thulasiraman, N. Swamy, Graphs: Theory and Algorithms, Wiley, 2011.

[10] J. Balthrop, S. Forrest, M.E.J. Newman, M.M. Williamson, Technological networks and the spread of computer viruses, Science 304 (5670) (2004) 527–529, http://dx.doi.org/10.1126/science.1095845 (URL http://www.sciencemag.org/content/304/5670/527.short).

[11] E.-A. Horvát, M. Hanselmann, F.A. Hamprecht, K.A. Zweig, One plus one makes three (for social networks), PLoS One 7 (4) (2012) e34740, http://dx.doi.org/10.1371/journal.pone.0034740(URL http://dx.doi.org/10.1371%2Fjournal.pone.0034740).

[12] E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nat. Rev. Neurosci. 10 (3) (2009) 186–198, http://dx.doi.org/10.1038/nrn2575(URL http://dx.doi.org/10.1038/nrn2575).

[13] L.A. Zager, G.C. Verghese, Graph similarity scoring and matching, Appl. Math. Lett. 21 (1) (2008) 86–94, http://dx.doi.org/10.1016/j.aml.2007.01.006(URL http://www.sciencedirect.com/science/article/pii/S0893965907001012).

[14] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. del la Société Vaudoise des Sciences Naturelles 37 (1901) 547–579.

[15] Y. Wang, J.C. Fisher, R. Mathew, L. Ou, S. Otieno, J. Sublet, L. Xiao, J. Chen, M.F. Roussel, R.W. Kriwacki, Intrinsic disorder mediates the diverse regulatory functions of the Cdk inhibitor p21, Nat. Chem. Biol. 7 (2011) 214–221, http://dx.doi.org/10.1038/nchembio.536(URL http://dx.doi.org/10.1038/nchembio.536).

[16] G. Carlsson, Topology and data, Bull. Am. Math. Soc. (N.S.) 46 (2) (2009) 255–308, http://dx.doi.org/10.1090/S0273-0979-09-01249-X(URL http://dx.doi.org/10.1090/S0273-0979-09-01249-X).

[17] H. Edelsbrunner, J. Harer, Computational Topology — An Introduction, American Mathematical Society, 2010. (http://www.ams.org/bookstore-getitem/item=MBK-69).

[18] R. Ghrist, Barcodes: the persistent topology of data, Bull. Am. Math. Soc. 45 (2008) 61–75, http://dx.doi.org/10.1090/S0273-0979-07-01191-3(URL http://www.ams.org/bull/2008-45-01/S0273-0979-07-01191-3/).

[19] R.T. Rockafellar, R.J.B. Wets, Set convergence, Variational Analysis, Grundlehren der mathematischen Wissenschaften, vol. 317, Springer, Berlin Heidelberg, 1998, pp. 108–147, http://dx.doi.org/10.1007/978-3-642-02431-3_4, (URL http://dx.doi.org/10.1007/978-3-642-02431-3_4).

[20] V. Nanda, Perseus: The Persistent Homology Software, (Date 15.10.2012), URL http://www.math.rutgers.edu/vidit/perseus.html2012.

[21] W.-L. Hung, M.-S. Yang, Similarity measures of intuitionistic fuzzy sets based on Hausdorff distance, Pattern Recogn. Lett. 25 (14) (2004) 1603–1611.