BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 784–791
doi:10.1093/bioinformatics/btt618

Structural bioinformatics

Advance Access publication October 27, 2013

Fast protein fragment similarity scoring using a Binet–Cauchy kernel

Frédéric Guyon[1,*] and Pierre Tufféry[1,2,3]

[1]Univ Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques in Silico, UMR 973, F-75205 Paris, France, [2]INSERM, U973, F-75205 Paris, France and [3]Univ Paris Diderot, Ressources Parisiennes de Bioinformatique Structurale, F-75205 Paris, France

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Meaningful scores to assess protein structure similarity are essential to decipher protein structure and sequence evolution. The mining of the increasing number of protein structures requires fast and accurate similarity measures with statistical significance. Whereas numerous approaches have been proposed for protein domains as a whole, the focus is progressively moving to a more local level of structure analysis for which similarity measurement still remains without any satisfactory answer.

Results: We introduce a new score based on Binet–Cauchy kernel. It is normalized and bounded between 1—maximal similarity that implies exactly the same conformations for protein fragments—and −1—mirror image conformations, the unrelated conformations having a null mean score. This allows for the search of both similar and mirror conformations. In addition, such score addresses two major issue of the widely used root mean square deviation (RMSD). First, it achieves length independent statistics even for short fragments. Second, it shows better performance in the discrimination of medium range RMSD values. Being simpler and faster to compute than the RMSD, it also provides the means for large-scale mining of protein structures.

Availability and implementation: The computer software implementing the score is available at http://bioserv.rpbs.univ-paris-diderot.fr/BCscore/

Contact: frederic.guyon@univ-paris-diderot.fr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 17, 2013; revised on October 18, 2013; accepted on October 22, 2013

1 INTRODUCTION

When analyzing biological processes, similarity measures and statistical significance are one of the major means we have to establish relationships and infer models underlying observations. In the field of protein structure, the matter of quantifying similarity between two structures is a long standing objective, as it is an essential key to decipher protein sequence–structure–function relationships and further to classify them.

Protein similarity search can be performed at a global and a local level. Whole structure comparisons provide general information about protein classification and protein functions. At a more local level, fragment comparison and identification has become a key step for protein structure analysis, annotation and modeling. Fragment similarities reveal functionally important residues (Tendulkar *et al.*, 2010), similar structural motifs may indicate function preservation in remote homologs (Manikandan *et al.*, 2008), and more generally, recurring fragments may be used as building blocks to the construction of *de novo* models of protein structures (Bystroff *et al.*, 1996; Friedberg and Godzik, 2005; Samson and Levitt, 2009; Unger *et al.*, 1989).

Scores used to quantify the 3D similarity are essential components at both global and local levels. A considerable amount of approaches has been developed for complete protein comparisons such as SSAP (Orengo and Taylor, 1996), DALI (Holm and Sander, 1995), CE (Shindyalov and Bourne, 1998), MAMMOTH (Ortiz *et al.*, 2002) or TM-align (Zhang and Skolnick, 2005) to cite some. However, progress in modeling and in the analysis of the significance of local differences in homologous proteins stresses the necessity to focus on smaller structure sizes. Yet none of the above scores used for structural alignment seems well adapted to the large-scale comparison of short protein fragments (Guyon and Tuffery, 2010). First, speed is of the essence for large-scale structure mining, and structural alignment may require heavy computations. Second, a crucial and difficult question remains that of the statistical significance of the scores applied to short fragments.

The root mean square deviation (RMSD) (Coutsias *et al.*, 2004; Kabsch, 1976, 1978) has been one of the first measures introduced. This long used criterion also has well-known flaws, among which, RMSD dependence on the alignment length, possibly large values between homologous proteins, and above all, a poor classification performance for medium range RMSD values: small RMSD values imply correct similarity, but large RMSD values can hardly be related or not to the absence of similarity. Numerous studies have attempted to overcome these limitations such as Betancourt and Skolnick (2001), Carugo and Pongor (2001), Chew *et al.* (1999), Kedem *et al.* (1999), Maiorov and Crippen (1995) and many others since. Among these, the unit-vector RMS distance (URMS) (Chew *et al.*, 1999; Kedem *et al.*, 1999) has been reported to be rather length independent and insensitive to local structural dissimilarities for sufficiently long structures. The TM-score (Zhang and Skolnick, 2004) has been designed to be rather insensitive to small structural deviations. It is also normalized to be length insensitive, but its design was clearly focusing on sufficiently large proteins (>80 amino acids).

*To whom correspondence should be addressed.

Here, we introduce a new, fast and accurate scoring scheme for fragment mining. Contrarily to the RMSD, fragment superimposition is not required. This new measure is based on a Binet–Cauchy kernel (BC score). This kernel has already been applied in unrelated context, in particular to the difficult problem of clustering video sequences and to the discrimination of an individual or a group of individuals in a video sequence (Wolf and Shashua, 2003; Vishwanathan and Smola, 2004).

In the context of protein structure comparison, the BC score can be seen as a shape similarity score corresponding to a correlation score between fragment shapes. Hence it is normalized and its values range from −1 measuring perfect shape anti-similarity (one fragment is the mirror image of the second one) to 1 indicating perfect similarity (up to a linear deformation). The BC score has several interesting properties. In particular, it is independent to any rotation of the structures and consequently its computation does not involve a prior superimposition of the structures. It is also fast and simple to compute—the score only requires the computation of 3×3 matrix determinants. Therefore, it is especially well adapted to perform large-scale protein mining and is designed to compare short protein fragments.

In this article, we simply assess a new gapless approach of structure similarity and we do not study the optimal structural alignment based on the BC score. After introducing the principle of the BC kernel, we review some of its important mathematical properties. We then present a statistical analysis of the scores and we finally assess the sensitivity and specificity in the context of large-scale fragment mining with a comparison with the RMSD. We also apply the BC score to symmetry detection in proteins.

2 METHODS

2.1 RMSD calculation and Kabsch formula

To introduce mathematical notations, we briefly recall well-known facts about RMSD calculation. Given two sets of aligned atoms, their superimposition consists in computing the optimal rigid body translation and rotation, which moves one set of atoms onto the second one minimizing the sum of squared Euclidean distances between them.

The two sets of atom coordinates are represented by two $N \times 3$ matrices X and Y, with X_{ij} (respectively, Y_{ij}) denoting the j^{th} coordinate of the i^{th} atom. The coordinate deviation is $\sum_{j=1}^{N} \sum_{i=1}^{3} (X_{ij} - Y_{ij})^2$ or using the Frobenius norm of matrices $||X - Y||^2$. The first step of the superimposition consists in centering the two structures. After this translation, we have: $\sum_j X_{ij} = \sum_j Y_{ij} = 0$ for all $i, 1 \leq i \leq 3$. In the following, all coordinate matrices are centered. Next step is to find the rotation matrix that minimizes

$$\min_{R \text{ rotation matrix}} ||XR - Y||^2 \tag{1}$$

Mathematically, a rotation matrix is characterized by $R^\top R = Id$ and $\det(R) = 1$, and then the RMSD is given by

$$\text{RMSD}(X, Y) = \sqrt{\frac{1}{N} ||XR - Y||^2}$$

Many algorithms have been proposed in the past to solve this problem.

The solution of Kabsch (1976) and Kabsch (1978) can be formulated as follows:

$$\text{RMSD}^2(X, Y) = \frac{1}{N}(||X||^2 + ||Y||^2 - 2(s\sigma_1 + \sigma_2 + \sigma_3)) \tag{2}$$

where σ_i are the three singular values of the 3×3 matrix $X^\top Y$ with $0 \leq \sigma_1 \leq \sigma_2 \leq \sigma_3$ and s is the sign of the determinant of $X^\top Y$.

Interestingly enough, we show in this article that beside the sign of the determinant, its value itself is meaningful. We prove in the following that the det-function is a kernel function and can be used as a fast and accurate score to compare protein structures.

2.2 The Cauchy–Binet scores

We call a multi-index a m-tuple of m elements from $\{1, \ldots, N\}$. Multi-indices are ordered in lexicographic order: $S = (i_1, i_2, \ldots, i_m)$ with $1 \leq i_1 < i_2 < \ldots < i_m \leq N$.

The Binet–Cauchy identity states that for two matrices $X, Y \in \mathbb{R}^{n \times m}$, with $m \leq n$, we have

$$\det(X^\top Y) = \sum_{S/|S|=m} \det(X_S)\ \det(Y_S) \tag{3}$$

where the sum is over all multi-indices of m elements. X_S is the submatrix of X constructed by choosing columns of indices in S.

Suppose matrices X and Y are represented by the vector of $\binom{m}{n}$ components indexed by S called the Grassman vector:

$$\phi(X)_S = \det(X_S)$$

With such a representation, the Binet–Cauchy theorem prove that the determinant in (3) can be expressed as a inner product between $\phi(X)$ and $\phi(Y)$.

$$\det(X^\top Y) = \phi(X)^\top \phi(Y) \tag{4}$$

Otherwise stated, this proves that the measure $\det(X^\top Y)$ is a positive semidefinite kernel denoted $K(X, Y)$. The function ϕ maps X to a high dimensional vector space named feature space such that BC is the usual Euclidean scalar product in this space of dimension $\binom{N}{3}$.

As X refers to protein structures, space dimension is $m = 3$. Hence, Grassman vector $\phi(X)$ has $\binom{N}{3} = \frac{N(N-1)(N-2)}{6}$ components. Each component of $\phi(X)$ indexed by S is the signed volume of the parallelepiped formed/defined by the triplet S of atoms and point (0,0,0). This volume is six times the volume of tetrahedron formed by the same four vertices (Fig. 1). The size of the Grassman vector increases rapidly with the length of the fragment it represents (for N = 20, we have $\binom{N}{3} = 1140$ and $\binom{60}{3} = 34220$. The BC kernel permits to compute a scalar product

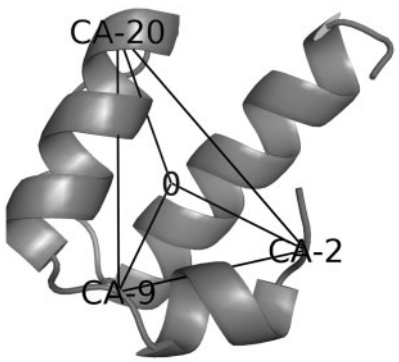

Fig. 1. One of the $\binom{46}{3} = 15180$ tetrahedrons of the domain d1a62a1 built from C-α of residues number 2, 9, 22 and center of all C-α O. Its (signed) volume is given by $\frac{1}{6}\det(X)$ where X is a 3×3 matrix containing the x,y,z coordinates of the three C-α

between such big vectors with a small amount of basic operations as it reduces to the calculation of a 3×3 matrix determinant.

2.3 Structural BC score derived from BC kernel

We only consider the coordinates of the α-carbon atoms of the protein fragments. The coordinates of the N residue fragments to be compared are stored in $N \times 3$ matrices X and Y. The coordinate matrices are centered at the origin. We propose a structural score derived from the Binet–Cauchy kernel. This score, we named Binet–Cauchy score is the cosine between the Grassman vectors of X and Y

$$\mathrm{BC}(X, Y) = \frac{\det(X^\top Y)}{\sqrt{\det(X^\top X)\det(Y^\top Y)}} \quad (5)$$

The calculation is possible only if the numerator is not equal to 0. This occurs when one of two matrices X or Y is not full rank, which is equivalent to state that one of the two fragments is absolutely flat. This never occurs in practice with protein fragments.

A second normalization of the BC kernel can also be considered:

$$\mathrm{BC}_2(X, Y) = \frac{2\det(X^\top Y)}{\det(X^\top X) + \det(Y^\top Y)} \quad (6)$$

As the geometric mean is always less than the arithmetic mean, we always have $\mathrm{BC}(X, Y) \leq \mathrm{BC}_2(X, Y)$. In the following, we focus on the first form of the BC score.

2.4 Mathematical properties of the Binet–Cauchy kernel

Here, we list some simple but important properties of the BC kernels.

2.4.1 The BC score is rotation independent For a rotation matrix R, we have $\det(R) = 1$. From the definition, it is clear that

$$\mathrm{BC}(X, YR) = \mathrm{BC}(X, Y)$$

It means that it is not necessary to compute a rotation matrix to optimally superimpose the two structures and compute the score.

2.4.2 The BC score is a correlation coefficient

$$\mathrm{BC}(X, Y) = \frac{\phi(X)^\top \phi(Y)}{||\phi(X)||.||\phi(Y)||} \quad (7)$$

The BC score is a cosine or a Pearson's correlation between the Grassman representation of X and Y. Therefore, we have

$$-1 \leq \mathrm{BC}(X, Y) \leq 1 \quad (8)$$

2.4.3 The BC score is a flexible score The BC scores are maximum for identical structures.

$$\mathrm{RMSD}(X, Y) = 0 \Rightarrow \mathrm{BC}(X, Y) = 1 \quad (9)$$

If X and Y can be exactly superimposed, then it exists a rotation matrix such that $Y = XR$. Because $\det(R) = 1$, we have $\mathrm{BC}(X, Y) = 1$. However, the reciprocal statement is not true. It is possible that $\mathrm{BC}(X, Y) = 1$ for two different fragment conformations with $\mathrm{RMSD}(X, Y) > 0$. The BC score is independent to any linear transformation with a positive determinant.

$$\mathrm{BC}(X, YA) = \mathrm{BC}(X, Y) \quad (10)$$

In particular, $\mathrm{BC}(X, XA) = \mathrm{BC}(X, X) = 1$ where A is a 3×3 matrix A with $\det(A) > 0$.

Nonetheless, in large mining experiments, the BC score shows weak sensitivity to protein fragment deformations. Therefore, it can be efficiently used to search for fragments in structure databases with a certain amount of flexibility. In next paragraph, we describe a mean to control that amount of flexiblity (See mathematical details in Supplementary Materials).

2.5 Tuning flexibility with distance constraints

We denote X_i (respectively, Y_i) the coordinates of the i^{th} (respectively, j^{th}) C_α of the fragment X (respectively, Y) of length N. The rate of deformation between the two structures is

$$\mathrm{defR}(X, Y) = \max_{1 \leq i \leq N} \frac{||X_i|| - ||Y_i||}{||X_i|| + ||Y_i||}$$

The rate of deformation compares the maximum variation of intra-distances between two residues distant of l residue.

It is also possible to reinforce the control over the opening of the fragment with

$$\mathrm{defR}'(X, Y) = \max\left\{\mathrm{defR}(X, Y), \frac{||X_N - X_1|| - ||Y_N - Y_1||}{||X_N - X_1|| + ||Y_N - Y_1||}\right\}$$

It can be verified that if $\mathrm{BC}(X, Y) = 1$ and $\mathrm{defR}(X, Y) = 0$ then X can be exactly superimposed on Y and $\mathrm{RMSD}(X, Y) = 0$.

The rate of deformation is used to control the admissible deformation between X and Y for a given value of BC.

2.6 BC distribution and Pareto distribution approximation

The distribution of the BC scores between two random fragments is equivalent to the distribution of the determinant of a random matrix. Except in special cases (distribution of a normal random matrices), it is difficult to derive a closed form of this distribution. Nevertheless, an accurate approximation of the tail of the distribution function is provided by the Pickands–Balkema–de Haan theorem (Balkema and de Haan, 1974; Pickands, 1975). The distribution of largest values of similarity scores, for example alignment scores, is widely used to assess the score significance for protein sequence or structure mining. The usual approach is based on extreme value distributions given by the Fisher–Tippett–Gnedenko theorem. Here, we are more interested in all the values above a given threshold and not only on the maximal value of the score. The reason is that considering only the best score, we lose a lot of information given by all the good scores. A second theorem in extreme value theory, named Pickands–Balkema–de Haan theorem gives the distribution of values above a given threshold called the exceedances. This theorem says that for a large class of score distribution, for a sufficiently high threshold, the distribution of the exceedances approximately follows with a good approximation a generalized Pareto distribution. To determine theoretical statistical properties of the BC scores, we generated random structures. In Shibuya (2010), a freely jointed chain model is used to assess complexity of fast RMSD searches of proteins structures. This model is random walk model where each step between two successive atoms has a fixed length and is a random vector independent from all other steps. This simple model is used to model polymers (De Gennes, 1979). To have a more realistic representation of structural fragments, we generated each random fragments from a real fragment by randomly permuting its inter C_α intervals and reconstructing the fragment as a random walk. Hence, the inter C_α distances and angle distribution corresponds to real fragment structures. Depending on the selected structures and on the fragment lengths, the number of comparisons could exceed 10^7 per fragment length. For each length, 10^5 fragment comparisons have been randomly sampled and their distribution parameters have been estimated. Parameter averages and standard deviations over 10^3 samplings have been computed and appear stable relatively to the sampling process (not shown). The distribution function depends on two parameters: the scale σ and shape k. The fit of a generalized Pareto model to the observed BC scores has been obtained with the R package evir using a maximum likelihood method (Pfaff and McNeil, 2012).

2.7 Data

All computations have been performed on protein domains provided by Astral-1.75 database with <70% sequence identity (release June 2009).

To derive statistics about the BC score distribution, a representative-selected subset of 64 Astral-1.75 domains have been randomly sampled consisting of 16 structures per main SCOP classes (alpha, beta, alpha/beta and alpha+beta proteins). An all-against-all fragment comparison has been performed for fragment of 20, 30, 40, 50 and 60 residues, and the BC deformation scores and RMSD have been computed for each fragment pair.

Finally, CASP10 target domains and models were downloaded from http://www.predictioncenter.org/casp10/. We used target/model pairs for 36 domains for which the experimental model has no missing residue, 13 free modeling and 23 template-based modeling (TBM and TMB-hard) domains. Only the models generated by servers were considered. We calculated the corresponding GDT_TS and TM-scores using the TMscore program (Zhang and Skolnick, 2004).

3 RESULTS

3.1 Comparison of RMSD and BC scores

Figure 2 represents the RMSD values versus the BC scores obtained from the comparisons of a fragment d3hxva_ (residues 5–25) against a representative subset of 1158 domains of Astral-1.75. This subset consists in three structures per SCOP family. Only family from classes a, b, c and d (alpha, beta, alpha/beta and alpha+beta proteins) are considered.

BC scores and RMSD are not well correlated with a Pearson correlation ρ equals to -0.146. The correlation is, however, stronger for low RMSDs. For instance, for RMSD ≤ 3, $\rho = -0.7$ and all pairs with RMSD <2 Å show a BC score over 0.6. However, the reverse is not true. Pairs of fragments with high BC scores present a large range of RMSDs from low to relatively high RMSDs (>4 Å). Remember the BC score is invariant under linear transformation, and thus a measure of the amount of fragment flexibility induced by linear or nearly linear transformations [or distorsion score - see Equation (10)] is necessary. Using a maximum deformation score of 0.4—black dots of Figure 2—one observes a shift of the distribution toward low RMSDs, mostly for BC scores in the medium range of values. However, this bracketing of the deformation still corresponds to rather large RMSD, even for BC scores of 0.8.

Strong BC scores indicate that the global fragment shapes are conserved but can be locally distorted (case B compared with A). For instance, looking at fragments B and C, B presents a much higher BC score for an approximately equivalent RMSD. The reason is that its shape is better correlated to the query as both fragment ends can be perfectly aligned with the query terminus.

Lower absolute values of the BC score imply medium to high RMSDs. They indicate a lack of shape similarity (case F) even for low RMSD values (case E). Finally, negative scores, close to -1, occur for mirror similarities, when one fragment is close to the mirror image of the second one (case G). The fragments corresponding (case G) have been superimposed by RMSD minimization and yet do not properly render the mirror symmetry. In that example, the helix part is shifted by a half helix turn relatively to the other one. When the two helix parts are superimposed, the fragments are then rotated by 90 degrees and the symmetry becomes clear (Fig. 2G). Interestingly, the occurrences for largely negative values are much less than for positive values, indicating the mirror conformations are much less observed in protein structures.

Also, the examples D, E, F and G illustrate that medium-range RMSDs do not imply significant conformation similarity. On the contrary, the BC score more precisely characterizes global shape similarity and combined with distortion rate allows for discriminating between spurious and true fragment 3D similarity.

3.2 Comparison of DALI and distance deviation scores versus BC scores

We also performed large-scale comparison based on the DALI elastic similarity score (Holm and Sander, 1995), which is a well known and used score based on distance matrices of the fragments. As shown in Figure 2, relatively to the BC score, the DALI score and the RMSD tend to behave in a similar way: they are insensitive to local deformations and mostly sensitive to global one (see points A, B, G). Internal distances are

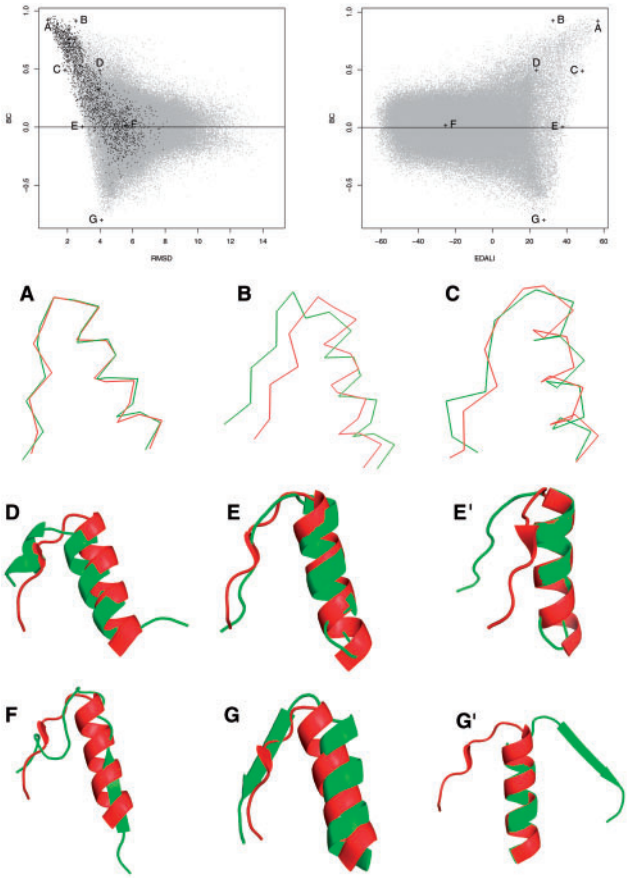

Fig. 2. Representation of 145 000 fragment comparisons of d3hxva_ Astral entry residues 5–25 against the 64 protein representative set. Top plots represent BC score versus the RMSD and the EDALI score. Query:d3hxva_ from 5 to 25; (**A**) d1wr8a_ from 21 to 41, rmsd = 0.79, BC = 0.92, defR = 0.057; (**B**) d1ovma1 from 15 to 35, rmsd = 2.52, BC = 0.91, defR = 0.50; (**C**) d1o0xa_ from 38 to 58, rmsd = 1.85, BC = 0.49, defR = 0.33; (**D**) d1s3sg_ from 63 to 83, rmsd = 3.98, BC = 0.49, defR = 0.50; (**E**,**E′**) d1zzka1 from 20 to 40, rmsd = 2.91, BC = 0.0024, defR = 0.50; (F) d2o8la1 from 93 to 113, rmsd=5.61, BC = 0.015, defR = 0.72; (**G**,**G′**) d16pka_ from 316 to 336, rmsd = 4.092, BC = −0.80, defR = 0.7

insensitive to mirror configurations as distances between residues are preserved in reflected images. Therefore, by construction the DALI score is blind to mirror transformations. Such local mirror transformations can be observed at a local level where they indicate opposed side-chain orientations, which may imply important changes for protein function. Similar results—not shown—could be observed for the rate of deformation defR that is also a measure based on intra-distance distortions. On the contrary, the BC is sensitive to small deviations causing local mirror configurations as they change determinant signs in the BC formulation. The BC score is not changed by linear transformations. Even if real protein fragment distortions are not linear, it is less sensitive to large deviations implied by fragment stretching and bending.

3.3 BC scoring of CASP10 models

To further illustrate BC scoring properties, we have also investigated its behavior for comparing CASP10 models with the experimental conformations. Figure 3A and B show average scores between all the predicted models submitted for CASP10 and their corresponding target. Overall, the BC score is well correlated with the TM-score and GDT_TS with a Pearson correlation score of 0.93 and 0.86. Figure 3 depicts more in detail the relationship between the BC-score and the TM-score. For sake of clarity, the data is only plotted for all the models of 10 TBM domains. It is clear that high BC scores correlate perfectly with high TM scores (Fig. 3A). Differences in scoring arise for medium range values. As examples, three models (Fig. 3 labels 2, 3, 4) of target T0756d1 have a good identical TM-score (TM = 0.71) but present a decreasing BC score from high (BC = 0.812) to non-significant (BC = 0.143). Here, the BC score is in agreement with an increasing RMSD from 3.72 to 9.40 Å. The reason is that the TM-score is not penalized by locally discordant regions which in that case deteriorate the global shape of the model. In that case, poor modeling of the extremities of the two models B and C explained the higher RMSD and lower BC score. The same conclusions stand for the comparison of the GDT_TS and BC score. The BC score is a measure of global shape similarity. Localized structural discrepancies may deteriorate the scoring of a good model presenting a strong accuracy in important region of the protein. This shows that the BC score applied on complete models is not sufficient to assess the quality of a protein structural model, but its adaptation for model comparison is out of the scope of the present study. On the contrary, models with a correct global shape but with local structure mismatches can have a significant BC score with lower TM scores or GDT_TS as BC is insensitive to local deformations. Cases depicted with labels 5, 6 in Figure 3 presents such a situation where the two models with an identical BC score (0.70) have a high and a medium (TM = 0.74 and TM = 0.46) TM-score. The two models present local differences revealed by the TM-score (illustrated in Supplementary Material Fig. S1).

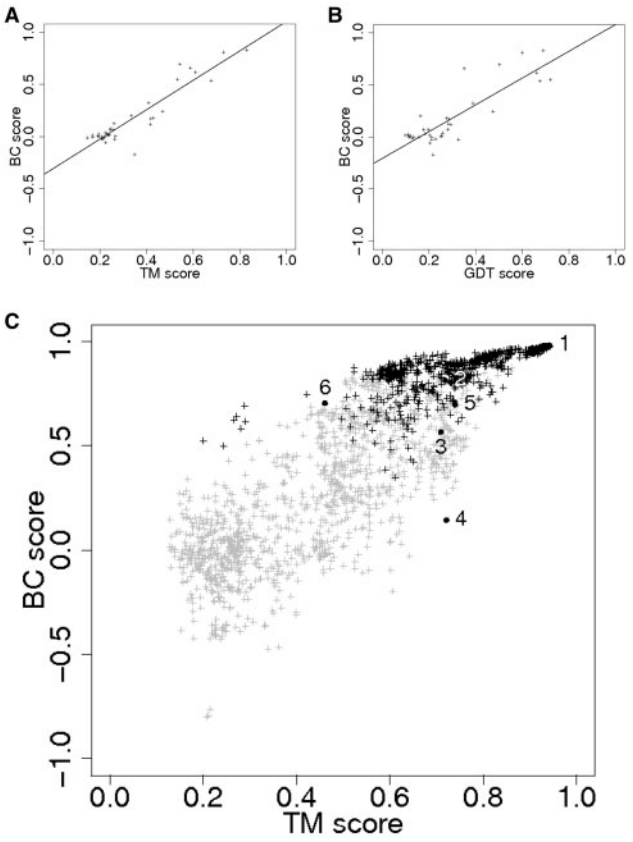

Fig. 3. (**A** and **B**): Correlation of BC score with TM-score (A) and GDT_TS (B). Plots show mean scores of the submitted models for each CASP10 targets. (**C**) Plot of scores for the models of 10 TBM targets. Dark points correspond to hits with RMSD <5 Å. 1: T0664d1/PconsM_TS4, rmsd = 2.549, BC = 0.978, 0.9424, Len = 498; 2: T0756d1/UGACSBL_TS1, rmsd = 3.72, BC = 0.81, TM = 0.71, Len = 91; 3: T0756d1/Phyre2_A_TS3, rmsd = 5.32, BC = 0.56, TM = 0.70, Len = 91; 4: T0756d1/AOBA-server_TS2, rmsd = 9.39, BC = 0.14, TM = 0.72, Len = 91; 5: T0743d1/BAKER-ROSETTASERVER_TS5, rmsd = 4.51, BC = 0.69, TM = 0.73, Len = 114; 6: T0743d1/Zhang-Server_TS2, rmsd = 5.99, BC = 0.70, TM = 0.46, Len = 114

3.4 Distribution of simulated fragments for different lengths

We now turn to the distribution of BC scores. Figure 4 and Supplementary Table S1 report the distributions obtained for random fragments obtained by a chain model, and for real fragments with no rearrangements. For random fragments (Fig. 4A), BC score densities are almost exactly identical and independent of the fragment lengths. Considering real fragments with no rearrangements (Fig. 4B) one observes density shapes vary according to fragment length. However, the average and standard deviations remain remarkably independent on fragment length and the density tails and the exceedance probability distributions modeled by a generalized Pareto distribution also converge rapidly to a limit distribution with parameters almost independent of the fragment length (Supplementary Material Table S2).

Such length independent behavior can be compared with that of the RMSD. On the exact same set, RMSD distribution (Fig. 4C) is clearly varying depending on fragment length. Another structural distance—URMS—considering 'difference' vectors between successive α-carbon instead of the position vectors of atoms has been proposed by Chew *et al.* (1999) and Kedem *et al.* (1999). It shows a better length invariance property compared with RMSD. For sufficiently long structures, the URMS is length independent and rather insensitive to local

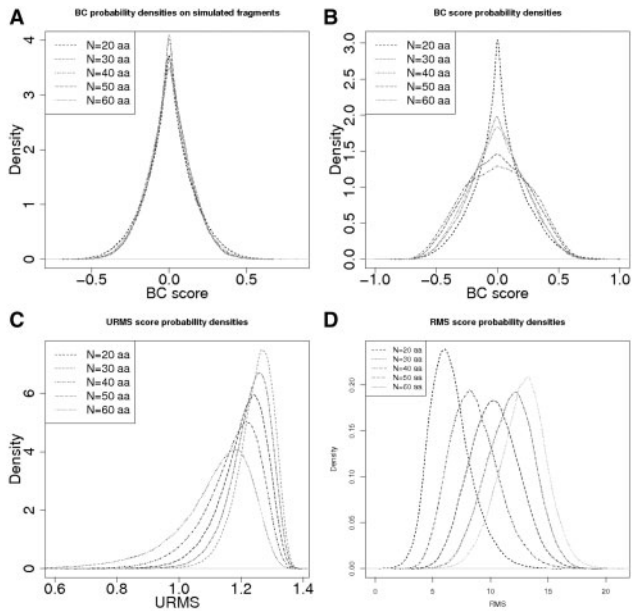

Fig. 4. Probability densities of BC score, RMSD and URMS for different fragment lengths (from 20 aa to 60 aa). These densities are computed over random fragments following the chain model (**A**) and over all real fragments of the representative set of 64 protein domains from Astral-1.75 databank (**B–D**)

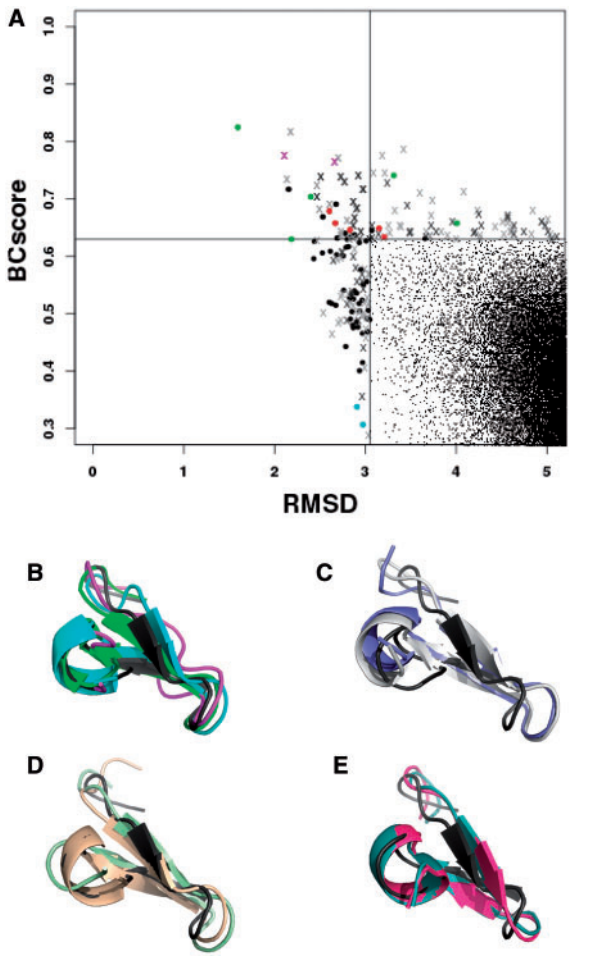

Fig. 5. Search for fragments similar to the tachycitin chitin binding motif (PDB:1dqc, residues 40–60). (**A**) Matches having RMSD <3.05 Å or BC score >0.63 are detailed. Dots, dark crosses, light crosses: matches having defR values <0.3, between 0.3 and 0.4, >0.4, respectively. Green and cyan dots: structures depicted in (**B**, **C** and **E**), respectively. Magenta crosses: structures depicted in (**D**). Red dots: structures annotated with the chitin key word in Uniprot. B–E: The query structure PDB 1dqc:40–60 is depicted in dark gray. (B) Structures having both low RMSD (<3.05) and large BC score (>0.63). (C) Structures having both large RMSD (>3.05) and BC score values. (D) Structures having large BC scores and large defR values (>0.4, magenta crosses in A). (E) Structures having both low RMSD and low BC score values

structural dissimilarities. However, probability density convergence appears in practice for structure length >80 residues, which is sufficient for comparing whole protein models but not for short fragments, for which the BC score appears much more stable.

3.5 Test cases

We further illustrate the behavior of the BC score in two different contexts.

3.5.1 Chitin binding domains We first illustrate the behavior of the BC score by mining a chitin binding motif, which is a functional motif found across plants, insects and animals. Suetake *et al.* (2000) have highlighted a structural similarity between motifs in tachycitin, an invertebrate chitin binding protein (PDB:1dqc, residues 40–60) and hevein, a protein of the rubber-tree (PDB 1hev, residues 12–32). Using the HHsuite sequence alignment facility (Remmert *et al.*, 2012), no similarity can be detected between the sequences of tachycitin and hevein.

The RMSD and the BC score between the two fragments are of 3.05 Å and 0.63, respectively. The value of 0.63 is associated with a *P*-value of 2.10^{-3}, i.e. a low significance. We have performed a search for similar fragments mining the Astral-1.75 subset at 70% sequence identity, not including 1dqc nor 1hev. The search using 1hev as seed happens to be an easy case, best hits having an RMSD <1 Å and BC scores >0.97. Searching fragments similar to 1dqc is more challenging and Figure 5 illustrates how the BC and defR scores can perform compared with the RMSD. We discuss only the matches having a BC score value >0.63 or a RMSD <3.05 Å. In all, 126 matches have an RMSD <3.05 Å from the 1dqc: 40–60 fragment. Their BC scores range from 0.3 to over 0.8. Figure 5B depicts three matches having both a low RMSD and a high BC score and Figure 5E two matches having a low RMSD and a lower BC score between 0.3 and 0.4. These matches mostly differ in their beta hairin conformation, which the BC score is able to discriminate. Conversely, 110 matches have a BC score of > 0.63. They include conformations having RMSD values up to 5 Å, and the additional use of the defR can be used to prune these matches. Figure 5B depicts two matches having a beta hairpin conformation similar to the one of 1dqc, but stretched. These two matches have an RMSD >3.05 Å (3.3 and 4 Å), a BC score >0.63 and a defR <0.3—i.e. a low deformation rate, which highlights the flexibility of the BC score. Figure 5D shows two

conformations of both large BC score and low RMSD values, for which the defR value is > 0.4. They have a beta hairpin conformation similar to that of the query, but differ in their extremities. Overall, such example illustrates that the combination BC score-defR allows a fine tuning of the conformational search. Interestingly, selecting the matches having a BC > 0.63 and a defR < 0.3 reduces the matches to only 17, among which 6—red dots, 2 are overlapping —have a chitin binding annotation in the Uniprot database (TheUniProtConsortium, 2012), and none of the matches rejected by the defR has such annotation.

We have also considered using the TM-score, which depends on several parameters that are tuned to assess longer structural alignments. Not surprisingly, as is, the use of TM-score for fragment mining is inadequate and results in a lack of precision. For instance, the TM-score between the two queries 1dqc and 1hev is 0.1982. Mining of the Astral-1.75 using TM-score with this value as threshold, 837 329 fragments has been retrieved.

3.5.2 Mirror conformations A second example illustrates the possibility to search for mirror conformations. Novotny and Kleywegt (2005) have shown that even though left-handed helices are rare, they occur in protein structures and are important for the stability of the protein, for ligand binding, or as part of the active site. When they do occur, they are structurally or functionally significant. We have performed the search for such left-handed helices over the collection of structures presented by Novotny and Kleywegt (2005), using as seed a right-handed helix fragment of Astral entry d1or4a at position 125. The left-handed helices identified by Novotny and Kleywegt (2005) have an average BC score of −0.96, with one outlier value of 0.81 for 1h21 chain A for which the left helix had a shorter shape. Interestingly, we also find that for one protein (PDB 1bnl chain A) another region had a propensity for left-handed helix with score on the same order, although not as helical. This region, although far in the sequence, is close to that displaying the left-handed helix previously identified (Fig. 6). Extending the search to larger collections, we could identify 269 and 368 structures displaying such left-handed propensities in the Astral 70 and PDB 70 collections, respectively. Some of them, such as notch1 and 2 had up to three left-handed helices.

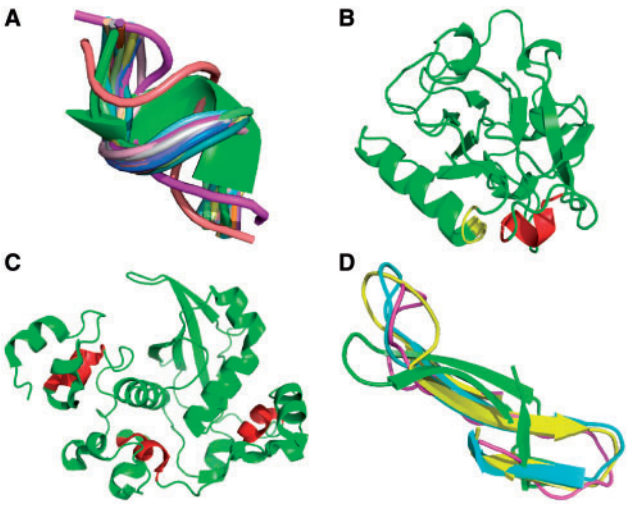

Fig. 6. Mirror similarities. (**A**) Left-handed fragments identified using a right-handed helical fragment as seed (green). (**B**) 1bnl entry. Red: left-handed helix identified in Novotny and Kleywegt (2005). Yellow: identified fragment with a largely negative BC score of −0.96. (**C**) Left-handed helices identified in the human notch1 protein (PDB:3eto). (**D**) WW motifs with a twist opposite to that of the 2zaj 17–40 fragment

Interestingly as well, the search for mirror conformations can also be performed for any shape. Figure 6D shows for instance the BC kernel can identify negative BC scoring fragments using as seed the three-stranded beta sheet of PDB 2zaj fragment 17–40. Note the scores of the three matches depicted (d1dyka2:41-64, d2b79a1:36-59, d1uaia_:114–137) are, however, on the order of −0.65 to −0.70, i.e. much less significant than those of left-handed helices.

These two examples highlight that the BC score can be used to quickly mine collections of structure to search for mirror conformations, such possibility has to our knowledge no equivalent so far.

4 COMPUTATIONAL COST

Finally, we have compared the execution times for mining fragments of length 10 using RMSD and BC score. The BC mining is >50 times faster than the same search using RMSD. To mine the complete Astral 1.75 subset at 70% sequence identity, typical execution times are on the order of 7 s on an Intel Xeon(R) CPU E5506 at 2.13GHz using four cores.

5 DISCUSSION AND CONCLUSIONS

In the present study, we have introduced the BC score, a new, fast and efficient scoring approach of protein fragment similarity. We have shown that its inherent geometric properties make it possible to control the fine tuning of the level of similarity, with the limit that being invariant on linear transformation, a control over the accepted deformation rate is required. Our tests show, however, that a typical maximal value of 0.3 for the deformation rate can be used as default. Accepting it, the BC score comes with two interesting features compared with the RMSD and its derivatives. First, its size independence property has so far only been addressed for long fragments. Our results show that the BC score reaches size independence even for short size (10–50 amino acids). This independence property allows for a simple estimation of a *P*-value based on exceedance value distribution. Second, we have also illustrated for remote homologs that the BC score makes it possible to identify fragments globally distorted, being not too sensitive to local structural changes. This results in the identification of relevant hits at larger RMSD. The generality of this behavior is, however, difficult to assess, although large-scale fragment comparisons such as that of Figure 2 clearly shows the BC score is able to assign high scores to fragments with large RMSD. Finally, the BC score also defines a meaning for the lower bound limit that corresponds to mirror conformations. As illustrated for left-handed helices, this makes simple the search for particular such conformations.

A major limit of the application presented here is its gapless property. In our experience, this can lead to sliding matches along a protein when performing systematic searches. Gap introduction, however, could be performed in the context of dynamics

programming even if the objective function derived from the BC score is not additive. These promising applications are the subject of further work. It remains that the BC score performs particularly well for gapless similarity search, and it is particularly fast compared with previously proposed scores because it does not require any superimposition, relying only on the calculation of 3×3 determinants. The weak computational cost of the BC score clearly opens the door to systematic large-scale analyses.

Funding: INSERM UMR-S973 recurrent funding and the BIP:BIP project ,"Investissement d'Avenir" ANR grant.

Conflict of Interest: none declared

REFERENCES

Balkema,A. and de Haan,L. (1974) Residual life time at great age. *Ann. Probab.*, **2**, 792–804.

Betancourt,M.R. and Skolnick,J. (2001) Universal similarity measure for comparing protein structures. *Biopolymers*, **59**, 305–309.

Bystroff,C. *et al.* (1996) Local sequence-structure correlations in proteins. *Curr. Opin. Biotechnol.*, **7**, 417–421.

Carugo,O. and Pongor,S. (2001) A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.*, **10**, 1470–1473.

Chew,L. *et al.* (1999) Fast detection of common geometric substructure in proteins. *J. Comput. Biol.*, **6**, 313–325.

Coutsias,E. *et al.* (2004) Using quaternions to calculate RMSD. *J. Comput. Chem.*, **25**, 1849–1857.

De Gennes,P.G. (1979) *Scaling Concepts in Polymer Physics*. Cornell University Press, Ithaca, NY.

Friedberg,I. and Godzik,A. (2005) Connecting the protein structure universe by using sparse recurring fragments. *Structure*, **13**, 1213–1224.

Guyon,F. and Tuffery,P. (2010) Assessing 3D scores for protein structure fragment mining. *Open Access Bioinformatics*, **2**, 67–77.

Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.

Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **32**, 922–923.

Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystall. A*, **34**, 827–828.

Kedem,K. *et al.* (1999) Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins*, **37**, 554–564.

Maiorov,V.N. and Crippen,G.M. (1995) Size-independent comparison of protein three-dimensional structures. *Proteins*, **22**, 273–283.

Manikandan,K. *et al.* (2008) Functionally important segments in proteins dissected using Gene Ontology and geometric clustering of peptide fragments. *Genome Biol.*, **9**, R52.

Novotny,M. and Kleywegt,G.J. (2005) A survey of left-handed helices in protein structures. *J. Mol. Biol.*, **347**, 231–241.

Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.

Ortiz,A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Pfaff,B. and McNeil,A. (2012) evir: Extreme Values in R. R package version 1.7-3. http://CRAN.R-project.org/package=evir (14 November 2013, date last accessed).

Pickands,J. (1975) Statistical inference using extreme order statistics. *Ann. Stat.*, **3**, 119–131.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Samson,A.O. and Levitt,M. (2009) Protein segment finder: an online search engine for segment motifs in the PDB. *Nucleic Acids Res.*, **37**, D224–D228.

Shibuya,T. (2010) Searching protein three-dimensional structures in faster than linear time. *J. Comput. Biol.*, **17**, 593–602.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension of the optimum path. *Protein Eng.*, **11**, 739–747.

Suetake,T. *et al.* (2000) Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. *J. Biol. Chem.*, **275**, 17929–17932.

Tendulkar,A.V. *et al.* (2010) FragKB: structural and literature annotation resource of conserved peptide fragments and residues. *PLoS One*, **5**, e9679.

TheUniProtConsortium. (2012) Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res.*, **40**, D71–D75.

Unger,R. *et al.* (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, **5**, 355–373.

Vishwanathan,S.V.N. and Smola,A.J. (2004) Binet-Cauchy kernels. In: *Proceedings of Neural Information Processing Systems NIPS'04*. Vancouver, Canada.

Wolf,L. and Shashua,A. (2003) Learning over sets using kernel principal angles. *J. Mach. Learn. Res.*, **4**, 913–931.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.