Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm

Spencer E Bliven[1,2,3,*], Aleix Lafita[1,4,5,*], Peter W Rose[6,7], Guido Capitani[1,8,†], Andreas Prlić[6,9], and Philip E Bourne[2,10]

[1] *Paul Scherrer Institute, Laboratory of Biomolecular Research, 5232 Villigen PSI, Switzerland*
[2] *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA*
[3] *Zurich University of Applied Science, Wdenswil, Switzerland*
[4] *Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland*
[5] *European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK*
[6] *RCSB Protein Data Bank, San Diego Supercomputing Center, UC San Diego, USA*
[7] *Structural Bioinformatics Laboratory, San Diego Supercomputer Center, UC San Diego, La Jolla, CA 92093, USA*
[8] *Department of Biology, ETH Zurich, 8093 Zurich, Switzerland*
[9] *Invitae*
[10] *Department of Biomedical Engineering, University of Virginia, 415 Lane Road, Charlottesville VA 22903*
[†] *Deceased*
[*] *Equal contributions. Corresponding authors: spencer.bliven@zhaw.ch, aleixlafita@ebi.ac.uk*

Abstract

Many proteins fold into highly regular and repetitive three dimensional structures. The analysis of structural patterns and repeated elements is fundamental to understand protein function and evolution. We present recent improvements to the *CE-Symm* tool for systematically detecting and analyzing the internal symmetry and structural repeats in proteins. In addition to the accurate detection of internal symmetry, the tool is now capable of i) reporting the type of symmetry, ii) identifying the smallest repeating unit, iii) describing the arrangement of repeats with transformation operations and symmetry axes, and iv) comparing the similarity of all the internal repeats at the residue level. *CE-Symm* 2.0 helps the user investigate proteins with a robust and intuitive sequence-to-structure analysis, with many applications in protein classification, functional annotation and evolutionary studies. We describe the algorithmic extensions of the method and demonstrate its applications to the study of interesting cases of protein evolution.

Availability: *CE-Symm* is an open source tool integrated into the BioJava library (`www.biojava.org`) and freely available at `https://github.com/rcsb/symmetry`.

Introduction

François Jacob described molecular evolution as a "tinkering" process, where pre-existing elements are combined and repurposed to solve new biological problems (Jacob, 1977). Traces of this "tinkerer evolution" can be seen in the widespread reuse of structural elements in proteins at different scales: small motifs (Lupas et al., 2001), functional domains (Han et al., 2007), and protein oligomerization (Levy and Teichmann, 2013). One example is the repetition of structural elements within a protein chain, thought to arise from gene fusion and duplication events (Andrade et al., 2001).

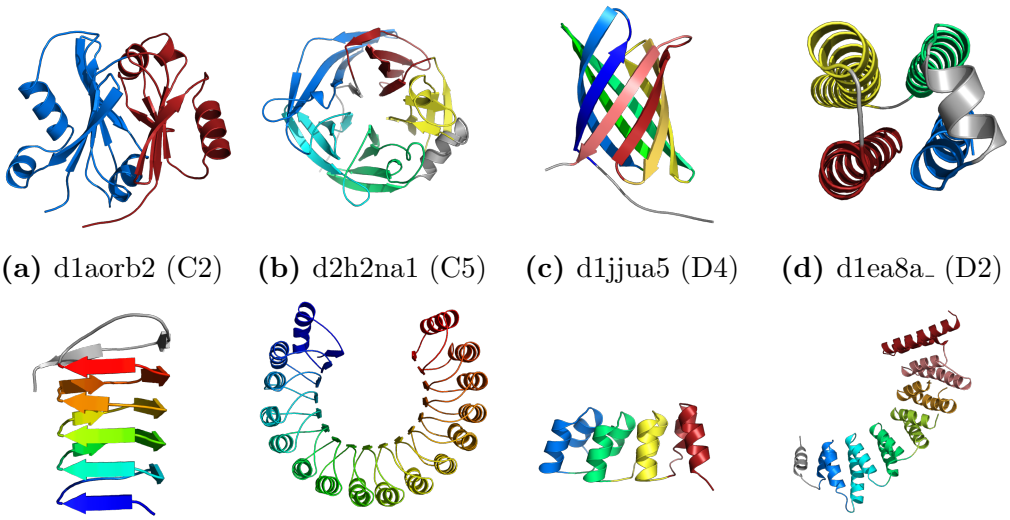
It is common for structural repeats in proteins to maintain a symmetric arrangement (Guerler et al., 2009), which has been associated with many biological functions (Goodsell and Olson, 2000). The internal symmetry of proteins is thought to arise from ancestral quaternary structures fused into a single polypeptide chain (Abraham et al., 2009; Lee and Blaber, 2011; Broom et al., 2012). However, since symmetric protein folds theoretically have a folding thermodynamic advantage, their symmetry could have arisen by evolutionary convergence (Wolynes, 1996). On the other hand, the evolution of functional patches is often symmetry breaking (Bonjack-Shterengartz and Avnir, 2015). High-quality alignments of structural repeats are essential to resolve these opposing evolutionary explanations and understand the tension between conservation and divergence.

A number of tools have been developed for the detection of structural repeats and internal symmetry in proteins (Kinoshita et al., 1999; Murray et al., 2004; Shih et al., 2006; Abraham et al., 2008; Guerler et al., 2009; Kim et al., 2010; Do Viet et al., 2015). Similarly to our *CE-Symm* method (Myers-Turnbull et al., 2014), the common approach to the problem is the identification of regions of similar structure within a protein chain, usually through the alignment of a protein structure to itself (self-alignment). However, the systematic extraction of repeating structural units and residue equivalencies among the repeats from the self-alignments is a nontrivial task. Here, we present an extension of *CE-Symm* (version 2.0) that accurately detects symmetry in proteins and defines the boundaries of the structural repeating elements. In addition, it reports the type of symmetry and describes the arrangement of repeats using the symmetry axes. Finally, the similarity of all the structural repeats can be compared at the residue level in a multiple structure alignment.

Types of symmetry

Several definitions of **internal symmetry** and **repeats** are possible, depending on the biological question of interest. For the purposes of this paper, we define it as the regular arrangement of a common repeating structural unit within a protein chain. Therefore, a repeat is an asymmetric structural motif present multiple times in the same structure. We restrict our consideration of repeats to cases where the orientation between adjacent structural units is regular; that is, where a consistent geometric transformation can be applied to superimpose each repeat onto the next. In other words, *CE-Symm* focuses on identifying repeats which conserve both the structure and the interface between repeats.

Several types of internal symmetry can be derived from this broad definition. The most basic division is between closed symmetry and open symmetry. In proteins with **closed symmetry**, the repeats are arranged in a point group symmetry. This can be defined mathematically as a set of rotations that superimpose equivalent repeats while keeping at least one point at the center of rotation fixed. In contrast, repeats in proteins with **open symmetry** are related by transformations with a translational component. Examples of closed and open symmetry can be found in Fig. 1a-d and Fig. 1e-h, respectively.



(a) d1aorb2 (C2) **(b)** d2h2na1 (C5) **(c)** d1jjua5 (D4) **(d)** d1ea8a_ (D2)

(e) d1l0sa_ (H3) **(f)** d1z7xw_ (R) **(g)** d1n0ra_ (R) **(h)** d1ltxa1 (SH)

Figure 1: Examples of protein domains with internal symmetry. Protein domains are labeled with SCOP domain identifier (Murzin et al., 1995). a) N-terminal domain of aldehyde ferredoxin oxidoreductase with 2-fold rotational symmetry (C2) of an alpha+beta motif; b) 5-bladed beta propeller with a helical insertion between second and third blades with 5-fold rotational symmetry (C5) of a 4-stranded beta-sheet motif; c) Beta-barrel with 8 beta strands in a 4-fold dihedral symmetry (D4) of single stranded motifs; d) 4-helical bundle with dihedral symmetry connectivity (D2) of single helical motifs; e) Beta-helix with single stranded right-handed helical symmetry (H3); f) Leucine rich repeats with open rotational symmetry (R) of 16 up and down alpha-beta motifs; g) Designed Ankyrin repeat protein with 4 translational repeats (R) of double helical motifs; h) Alpha-alpha right-handed superhelix (SH) of double helical motifs. Repeats are colored from blue, N-terminal, to red, C-terminal. Non-repeating parts of the structure are colored in grey.

Closed symmetries can be further characterized according to the possible chiral point groups: cyclic (Cn), generated by a single n-fold rotational operator (Fig. 1a-b); dihedral (Dn), which requires an n-fold rotation and n perpendicular 2-fold operators (Fig. 1c-d); and polyhedral point-groups (T, O, and I), which feature non-perpendicular rotation operators. Both cyclic and dihedral internal symmetries are common in proteins, but, although common at the quaternary structure level, polyhedral symmetries have not yet been observed within a single polypeptide chain.

Open symmetry can be further subdivided into special cases of helical, translational, and superhelical repeats. Helical symmetry consists of repeats arranged around a screw axis, where each repeat is related to the next by a fixed linear trans-

lation combined by a rotation around the central axis (Fig. 1e). In cases where the rotation angle is close to an fraction of a turn, we indicate the approximate number of subunits needed per turn (Hn). Proteins with open symmetry that have negligible translation are called rotational repeats (Fig. 1f), and those with negligible rotation between repeats are called translational repeats (Fig. 1g), both annotated as R. Superhelical symmetry (SH) provides the most general description of repeats with open symmetry, and is reserved for cases which cannot be expressed as a single fixed operator relating each repeat to the next. Instead, the rotation axis between adjacent repeats precesses along a helical path (Fig. 1h). Proteins with open symmetry are sometimes referred to as solenoid proteins (Kobe and Kajava, 2000).

Methods

CE-Symm analyzes the symmetry in a protein structure and produces a multiple alignment of all repeats, as well as ancillary information about the type and order of symmetry in the structure. An overview of *CE-Symm* alignment steps is shown in Fig. 2. These are described in detail below, but consist of (1) structural self-alignment, (2) order detection, (3) refinement to a multiple alignment, (4) Monte Carlo optimization of the multiple alignment, and (5) point group symmetry detection. These steps are repeated iteratively to detect multiple levels of symmetry (hierarchical symmetry) and higher-order point groups.

Self-alignment

CE-Symm begins with a structural self-alignment (other than the identity alignment) of the input protein structure using the Combinatorial Extension (CE) algorithm (Shindyalov and Bourne, 1998). Identifying significant self-alignments was the primary focus of the first version of the algorithm (Myers-Turnbull et al., 2014). In the self-alignment of structures with closed symmetry the first and last repeats are aligned, forming a circular permutation (CP) of the structure. This is why the structure alignment method used in *CE-Symm* shares algorithmic primitives with *CE-CP* (Bliven et al., 2015). For proteins with open symmetry, the initial self-alignment will always be missing one of the repeats due to the translation component of the symmetry operator.

The alignment quality is quantified using TM-Score (Zhang and Skolnick, 2004). Both inconsistently arranged repeats and large asymmetric regions in a structure will reduce the score of the self-alignment. In addition, open symmetry will generally have lower scores than closed symmetry, because the terminal repeats will be missing from the self-alignment.

Order of symmetry detection

The order of symmetry is defined as the number of symmetric units (repeats) in a structure. Extracting the order of symmetry is a key part of symmetry detection, and posterior analyses heavily depend on it. Several algorithms were evaluated for automatically determining the order (Bliven, 2015).

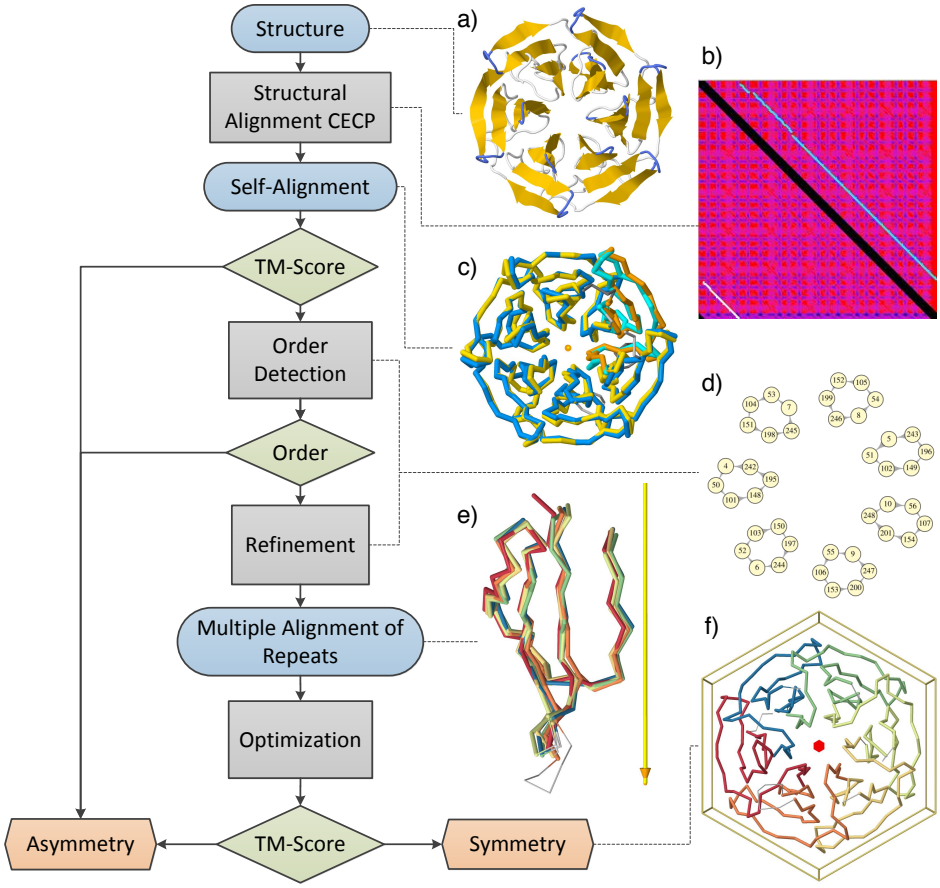


Figure 2: Flowchart of one iteration of the *CE-Symm* algorithm. Algorithm steps are grey rectangles, inputs and outputs are blue rounded rectangles, decision rules are green rhomboid boxes and final classifications are orange hexagonal rectangles. Additional iterations on the resulting repeats may be performed to detect further symmetry axes or hierarchical symmetry. The images on the right represent, from top to bottom: a) initial structure, colored by secondary structure elements; b) self-alignment dot-plot matrix, where similarity score is a range from blue (high similarity) to magenta (low similarity), the identity alignment is blacked out and the optimal self-alignment path is in white; c) superposition of the structure against itself based on the optimal self-alignment, where the original structure is in blue and cyan and a copy of the structure is in yellow and orange (orange and cyan correspond to the regions of the alignment involving a circular permutation); d) subset of the alignment graph with seven connected components of six aligned residues each; e) superposition of the six internally symmetric repeats according to the symmetry axis (yellow bar) and their residue equivalencies; and f) structure inside the six-fold cyclic symmetry (C6) box, with repeats colored differently.

The most straightforward method for determining the order in cases of closed symmetry is based on the angle of rotation, which can be calculated based on the superposition operator (Kim et al., 2010). The distance between a measured angle of rotation, θ, and the closest theoretical angle of rotation for order k is given by a triangle wave of frequency k:

$$\delta(\theta, k) = \frac{2\pi}{k} \left| \left(\left| \frac{\theta k}{2\pi} - \frac{1}{2} \right| \bmod 1 \right) - \frac{1}{2} \right| \tag{1}$$

The best-fit order is then the k that minimizes this distance, up to some maximum order.

Detecting the order based on the rotation angle works only for cases of closed symmetry. Another method, called graph-component order, can be used for both open and closed symmetry. Conceptually, the self-alignment is treated as a directed graph over the set of aligned residues. Residues that are aligned in all k repeats will form a path with k nodes. For open symmetry these paths tend to be disjoint, so simply finding the most frequent size of the connected components in the graph can accurately determine the order for open symmetry. For well-aligned cases of closed symmetry, the aligned residues form a cycle of k nodes, so the same method can also work in the general case. Those residues which participate in a path or cycle of the most frequent size form the refined alignment discussed in the following section.

For cases of closed symmetry, small alignment errors can lead to a situation where repeated applications of the alignment do not eventually form a closed cycle back to the original residue, but rather to a residue at a small offset. This can lead to failures of the graph-component order detector due to the merging of multiple alignment paths. This case can be handled by the sequence-function order detector. For each potential order k under consideration, all paths of length k are considered in that alignment graph. The number of residues separating the start and end of the path becomes a good indicator for the agreement between the graph and order k. Orders between 1 and some k_{max} (8 by default) are considered, with a pronounced decrease in average distance (40% by default) indicating that the correct order has been discovered. This method was already described in more detail in our previous publication (Myers-Turnbull et al., 2014).

By default, *CE-Symm* first runs the graph component order detector and, if the symmetry is determined to be closed, the sequence-function method is then used to improve the order determination.

Refinement to a multiple alignment

The refinement procedure takes as input the self-alignment of the structure and the order of symmetry and returns a multiple alignment of the repeats. *CE-Symm* has two implementations of the refinement procedure: graph-component and sequence-function, which are closely related to their respective order detectors.

The graph-component refiner combines the maximally connected components of the self-alignment graph with size equal to the order of symmetry. Each connected component contributes one column to the refined alignment of repeats, taking special care that the repeat sequences preserve the sequence order of the domain. A heuris-

tic method is used to decide which components to include in the refined multiple alignment.

The sequence-function refiner uses all the path of length k (the order of symmetry) to construct a multiple alignment of the symmetric repeats. The residues in each path are set as one column in the alignment, sorted in increasing order, resulting in a multiple alignment of the repeats. Note that, like the graph-component refiner, the multiple alignments obtained at the end of this stage do not contain gaps, so all repeats are of the same size.

Optimization

The multiple alignment obtained from the refinement is sometimes far from optimal, and depends very much on the goodness and consistency of the self-alignment. In addition, the refinement process prioritizes precision over coverage, which means that only the best residue equivalencies will be included, resulting in a shorter multiple alignment. The goal of the optimization is to increase the multiple alignment length while keeping the RMSD low. Furthermore, the optimization procedure can improve parts of the alignment that were not fully represented in the self-alignment, and thus not captured in the refinement result.

The optimization process uses a similar approach to the Combinatorial Extension Monte Carlo (CEMC) multiple structure alignment algorithm (Guda et al., 2001). The multiple alignment can be described by a matrix, where the rows represent aligned structures and the columns represent aligned positions (residue equivalencies). Rearranging and modifying the entries of the matrix results in changes of the multiple alignment. There are four possible moves (changes in the multiple alignment):

1. **Expand**: increase the alignment length by extending the boundary of a group of sequential residues, chosen randomly. This move requires the addition of an alignment column.

2. **Shrink**: decrease the alignment length by decreasing the boundary of a group of sequential residues, chosen randomly. This move requires a deletion of an alignment column.

3. **Shift**: move a group of sequential residues, chosen randomly, of one structure (row), chosen randomly, one position to the right or to the left.

4. **Insert gap**: delete one entry of the matrix, chosen randomly. This is equivalent to inserting a gap in one residue position (column) of one structure (row).

The insertion of gaps allows for partial repeat similarities in the alignment. All moves take into consideration that rows of the alignment occur sequentially in the protein sequence, so unaligned residues between repeats can be considered either at the end of a repeat or the beginning of the following one. In addition, the shrink and insert gap moves have been biased, so that the probability of choosing an alignment column or an equivalent residue, respectively, is proportional to the average inter-residue distance of the given column or the given residue, respectively.

A geometric distribution with parameter 0.5 is chosen to allocate the probability among alignment columns. A schematic representation of the steps and how they affect the multiple alignment is provided in supplementary figure S1.

After each optimization step, an alignment score is calculated. The score function to be optimized has also been smoothed with respect to the original CEMC score to remove discontinuities:

$$S = \sum_{i=0}^{N} \sum_{k=0}^{L} \left[\frac{C}{1 + \left(\frac{d_{ik}}{d_0}\right)^2} - A \right] - G \tag{2}$$

N is the number of structures (rows) in the alignment; L is the number of equivalent positions (columns) in the alignment, including gaps; C is the maximum score of an alignment position (by default set to 20); d_{ik} is the average distance from aligned residue k in structure i to all its equivalent residues; d_0 is the structural similarity function parameter, as defined by the TM-score (Zhang and Skolnick, 2004); A is the distance cutoff penalization, which shifts the function to negative values when the maximum allowed average distance of an aligned position (d_c) is higher than d_{ik}; and G is a linear gap penalty term. Calculation of A using a distance cutoff parameter d_c (by default set to 7Å) is straightforward from the condition that the score S has to be 0 when $d_{ik} = d_c$. The shape of the score function for different values of d_c is shown in supplementary figure S2.

The acceptance probability of a move is proportional to the score difference and decreases proportional to the number of optimization steps.

$$p = \left[\frac{C - \Delta S}{C\sqrt{m}} \right] \left(1 - \frac{m}{M} \right) \tag{3}$$

m is the current iteration number, ΔS is the change in alignment score and M is the maximum number of iterations. The maximum number of optimization iterations is proportional to the length of the domain, by default a hundred times the number of residues in the protein. Optimization finishes either because it reaches the maximum number of iterations or in the case that no moves are accepted for a fraction of the total number of iterations (by default M divided by 50).

Recursive symmetry detection

So far, the procedure described can only identify symmetry operations that require a single axis. However, some structures present symmetries represented with more than one axis. This is the case for point groups other than cyclic, like dihedral symmetry, or structures with more than one level of symmetry, what we define as hierarchical symmetries. Multiple *CE-Symm* iterations are run in a recursive manner, i.e. repeats found in previous rounds are recursively fed into the next run until a non-significant result (no symmetry) is found. The goal is to find all the significant symmetry levels of a structure.

At the end of an iteration, repeats are extracted from the internal symmetry result and one of them is chosen as the representative, by default the N-terminal

repeat. Results of successive iterations are merged by combining the symmetry axes and multiple alignments, generating a unique result for the query structure.

The recursive symmetry detection allows better order determination for difficult cases (e.g., TIM barrels), because usually fractions of the order of symmetry are initially found (e.g., 2-fold instead of 8-fold). Continuing the analysis recursively breaks the structure down to the true asymmetric repeating units (e.g., with three levels of symmetry: 2-fold, 4-fold and finally 8-fold).

Significance

There are three decision checkpoints in the algorithm flowchart in Fig. 2. The first significance criterion for a symmetry result is the self-alignment TM-score. Like in the previous version, the default threshold value is set to 0.4. The second significance criterion is the order of symmetry. A symmetric structure must have symmetry order greater than 1 and the refinement of the self-alignment into a multiple repeat alignment has to be successful. The third significance criterion is the average TM-score of the multiple alignment of repeats, defined as the average TM-score of all pairwise repeat alignments. The default threshold value for the average TM-score is set to 0.36, because a 10% decrease from the original TM-score is allowed after refinement due to the restrictive conditions imposed on it. In addition the number secondary structure elements (SSE) of the final asymmetric repeating unit is considered. If the the number of SSE of each repeat is lower than the threshold, the result will be considered non-significant. For many applications it may be desirable to exclude simple repeat units (e.g. helical bundle proteins), but these are included in *CE-Symm* analysis by default in order to find the highest possible symmetry in a structure.

Symmetry type determination

The recursive symmetry detection identifies a collection of symmetry axes that describe the arrangement of repeats in the query structure. In many cases, several of these axes can be combined to form higher-order symmetries. For example, a two-fold rotation axis can be combined with another orthogonal axis to form dihedral symmetry. Near-identical rotation axes can also be combined to form higher-order rotational symmetry.

To determine the point group symmetry, we build on the algorithm described by Levy et al. (2006). The symmetry axes can be found efficiently by first considering only the centroids of each repeat, since they must be in a symmetric configuration if the entire complex is symmetric. To find all possible symmetry axes, the centroids are rotated around axes that go through the centroid of the whole structure using an orientation grid search in quaternion space (Karney, 2007). For each orientation, the RMSD of the aligned centroids is calculated. If the centroids align within a threshold, then all Cα atoms are superimposed. The symmetry axis is then defined by the rotation matrix of this superposition. If the RMSD is less than a threshold value (i.e., 5 Å), the symmetry operation is considered valid. Since symmetry operations form a group, only a few are needed to complete the full point group

This procedure allows the combination of axes that have been considered separately by *CE-Symm*. The point group is included in the final symmetry output and displayed to the user as a polyhedron box around the protein structure.

Internal symmetry dataset

For evaluation purposes we used the manually curated dataset of 1,007 domains selected randomly from the set of *SCOP* superfamilies, introduced in our previous study (Myers-Turnbull et al., 2014). A small number of classifications were updated to be more consistent with the new symmetry definitions, specially for the cases of open symmetry. The updated version of the internal symmetry dataset (v2.0), together with the reasons of the modified annotations, is available at `https://github.com/rcsb/symmetry-benchmark` and summarized in supplementary table S1. An important note is that in the evaluation of the previous version the open symmetry cases in the benchmark were part of the asymmetric (negative) set, while they are part of the symmetric (positive) set in this evaluation.

Results

Method evaluation

In our previous article we compared the performance of *CE-Symm* against other internal symmetry detection methods. In the new version, the detection performance has only been affected by the additional order detection and alignment refinement steps. The ROC curves of both versions are very similar, with a slight reduction of the false positives in the new one (supplementary figure S3). At the default TM-score threshold values for result significance, the false positive (FP) rate has decreased from 5.5% to 2.5%, while the true positive (TP) rate has been reduced from 81% to 76% in the benchmarking dataset. The bottleneck in symmetry detection continues to be finding a significant self-alignment, as we previously suggested.

The different methods for order detection perform similarly for closed symmetry cases in the benchmark (supplementary table S2). The simpler graph-component method performs worse than the others, but it is the only one that can be used for open symmetries, while the sequence-function detector performs better than the rotation-angle method, particularly for difficult cases. The evaluation of the default order detection strategy of *CE-Symm* is shown in supplementary figure S4. The strategy achieves an overall 89% precision, although high order open symmetries are still an open challenge.

On average for symmetric entries in the benchmark where *CE-Symm* could find symmetry, the optimization step extended the repeat length by 43%, reduced the RMSD by 1.8% and increased the average TM-Score of the repeat alignment by 19.6%. Furthermore, using optimization an additional 23 cases (9% of the symmetric structures in the benchmark) were correctly identified as symmetric (209 with optimization, 186 without), which is a 12% improvement in symmetry detection. Because the highest scoring alignment of the simulation trajectory is taken as the result, optimization can only improve the initial alignment.

Sequence-structure analysis

This new version of *CE-Symm* is the first tool capable of presenting internal symmetry as a multiple structural alignment between repeats. This feature provides a direct correlation between sequence and structure and can be used for comparative and evolutionary analysis of a variety of protein folds and families.

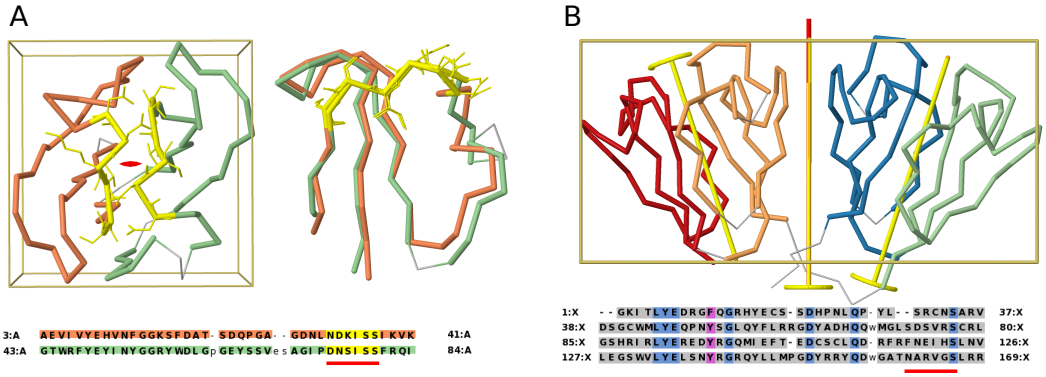


Figure 3: Internal symmetry in crystallin proteins A) An archaean $\beta\gamma$-crystallin with two repeats per chain (3HZ2). The full chain is displayed along its 2-fold axis, followed by a superposition of the repeats. The conserved calcium binding motif N/D-N/D-#-I-S/T-S is highlighted in yellow throughout. B) Human γ-D crystallin structure with four repeats per chain (1HK0). Two levels of symmetry exist: a C2 symmetry within each domain, and an additional C2 axis relating the two domains. The calcium binding motif has been lost (red bar below sequence), but other conserved positions (blue and magenta in the sequence) show the homology between the repeats.

Structural alignment of the repeats can reveal conserved motifs that have persisted since the duplication event. One example is the $\beta\gamma$-crystallin superfamily, which occurs in a variety of repeat arrangements. Many $\beta\gamma$-crystallins contain a calcium binding site motif (Aravind et al., 2009). As shown in Fig. 3A, the calcium binding motif is structurally conserved after a 2-fold rotation around the symmetry axis, and the residue side-chains preserve their orientation. Furthermore, calcium coordinates residues from both repeats, making the two-fold symmetry an essential feature of the binding site.

On the other hand, duplication events allow the appearance of asymmetry by independent sequence and structural divergence of the repeats. An example is the MaoC-like thioesterase/thiol ester dehydrase-isomerase superfamily (SCOP: d.38.1.4). Members of this family fold into a characteristic 'hot dog fold' which binds coenzyme A and catalyzes the dehydration of various bound fatty acids. Typically the MaoC-like proteins contain one hot dog domain per chain and assemble into dimers, tetramers, or hexamers (Pidugu et al., 2009). Some members of the family contain a duplication of the hot dog fold (Qin et al., 2000), accompanied by the loss of the catalytic motif R/N-####-H in one of the domains, in order to accommodate bulkier substrates which would otherwise not fit in a single domain (Pidugu et al., 2009). The structural divergence of the catalytic site in one of the repeats of the double hot dog subunit can be easily observed with *CE-Symm* (Fig. 4).

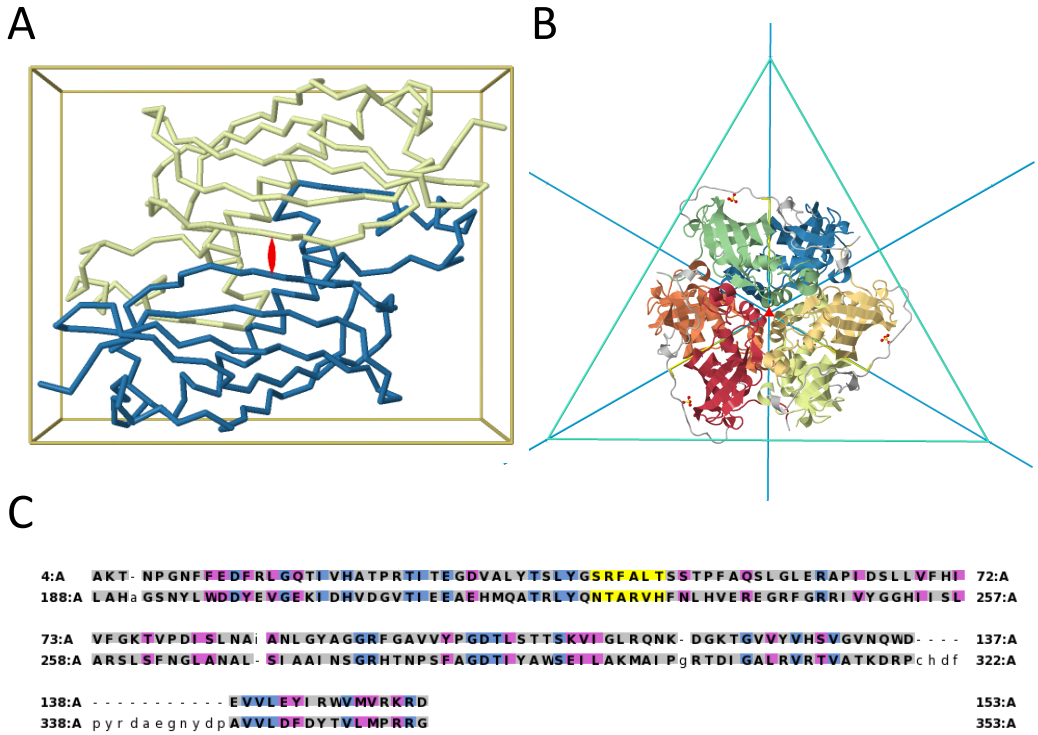


Figure 4: Hot dog fold duplication. A) Internal C2 symmetry in one chain of a MaoC domain protein dehydratase from Chloroflexus aurantiacus (4E3E) displaying a "double hot dog" fold. B) Full trimeric assembly, with the six individual hot dog domains colored. The quaternary structure has a threefold cyclic symmetry that combines with the twofold internal symmetry into a dihedral D3 symmetry equivalent to the quaternary structure of homologs without the internal domain duplication. C) Sequence alignment showing that the catalytic R/N-####-H motif (yellow) is lost in the first domain but retained in the second. Amino acid identity is shown in blue and similarity in magenta.

Multiple levels of symmetry

Some proteins contain more than one axis of symmetry. In those cases, the axes of symmetry can be collinear, orthogonal or independent to each other. If the axes are collinear, they can be combined into a single axis with higher symmetry order. If the axes are orthogonal, they can be combined into a point group of higher symmetry order.

If the axes are independent to each other, multiple levels of symmetry exist in the structure in a hierarchical organization. This can be an indication of multiple independent duplication events, like in the case of γ-crystallins (Fig. 3B), where four repeats are related by two independent 2-fold axes corresponding to two successive duplication events.

Internal symmetry and assembly stoichiometry

Additionally, the internal symmetry axes can also combine with the quaternary symmetry axes. Therefore, internal symmetry can increase the order of symmetry of a

protein complex. Returning to the previous MaoC-like protein example, the internal two-fold axis of the double hot dog domain in Fig. 4B is orthogonal to the three-fold quaternary symmetry axis, combining for an overall dihedral symmetry. This arrangement is structurally similar to the D3 quaternary symmetry of hexameric single hot dog proteins (e.g. 1YLI). Accounting for internal symmetry when comparing the symmetry two protein assemblies is therefore important, because proteins can have a similar overall structure despite their different subunit compositions. It would be misleading to say that the structure of the trimeric and hexameric MaoC-like proteins are substantially different. Another well-know example of similar overall arrangement with different subunit composition are DNA clamps, which promote processivity in DNA replication. In archaea and eukaryotes, the clamp is a trimer, while in bacteria it is a dimer (Kelman and O'Donnell, 1995). Furthermore, all DNA clamps have further internal symmetry axes leading to an overall D6 symmetry. As a historical note, the homology between bacterial and eurkaryotic DNA clamps was only acknowledged when the structures were solved and the similarity of their complexes was identified (Leipe et al., 1999).

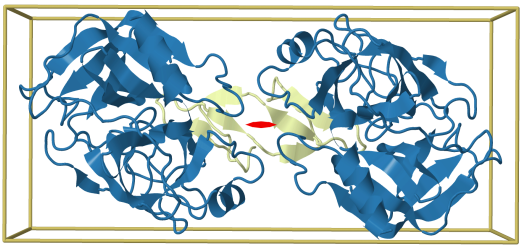


Figure 5: Uneven stoichiometry as a consequence of internal symmetry. The Bowman-Birk inhibitor from snail medic seeds (yellow) forms a complex with two bovine trypsin subunits (blue) in an uneven (A2B) stoichiometry ans asymmetric assembly (2ILN). However, a 2-fold symmetry axis (red) can be identified in the complex when the internal symmetry of the inhibitor is taken into account, showing that the complex is equivalent to one with even (A2B2) stoichiometry.

Furthermore, internal symmetry is important in understanding the stoichiometry of protein assemblies. Uneven stoichiometry assemblies are those with an unbalanced number of each entity type in the complex and occur rarely in the biological environment. It was previously reported that up to 40% of all protein assemblies with uneven stoichiometry in the PDB can be explained by the presence of internal symmetry in one or multiple of the subunits in the complex (Marsh et al., 2015). One such example is the artificial complex of Bowman-Birk inhibitor from snail medic seeds with bovine trypsin, which has an A2B stoichiometry (Fig. 5). Although the complex is asymmetric, considering the internal symmetry of the inhibitor shows that the assembly is structurally comparable to an even A2B2 assembly with C2 overall symmetry. This property has also functional consequences, since the binding of two trypsin proteins symmetrically allows the inhibitor to efficiently induce dimerization and block the peptidase activity. Symmetry is characteristic of biological assemblies and can be considered by methods, like EPPIC, in order to predict the biological assembly in the context of crystal latices (Bliven et al., 2017). Including internal symmetry in these methods could further improve predictions for some

known cases like, for example, uneven stoichiometries.

Open Symmetry

The majority of proteins with internal symmetry have closed symmetry. In the case of quaternary symmetry, this is expected since homooligomers with open symmetry are disfavored due to their aggregation potential (Capitani et al., 2015). However, this is not the case for open internal symmetry due to the ability for terminal repeats to diverge to avoid undesirable homotypic interactions. Open symmetry is common in internal repeats.

The most general formulation of open repeats in the literature is that of superhelical symmetry, where the repeated subunit is simultaneously translated along a helical path (curvature) and rotated around this path (twist) (Kobe and Kajava, 2000). *CE-Symm* cannot in principle identify superhelical symmetries, where both curvature and twist are relevant, because of the fundamental limit of the method to find a single symmetry axis (or multiple independent axes). However, we observed that the majority of structures containing tandem repeats that are classified as superhelical in the literature (solenoids) can be described by a single axis of symmetry. They fall in one of the following four conditions: i) the twist is negligible; ii) the curvature is negligible; iii) both twist and curvature are negligible; or iv) the twist is much larger than the curvature. In all those cases, *CE-Symm* can identify the symmetry in the structures and annotate them as helical, translational or open rotational symmetries.

For instance, from the 18 solenoid protein representatives from table 1 in Kobe and Kajava (2000), in 10 either the twist or the curvature are reported to be small (helical symmetry applies), in 5 both the twist and curvature are annotated as small (translational symmetry applies), and the remaining 3 structure representatives have irregular twist (asymmetric applies). Although many folds are classified as superhelical, only a small number have regular repeats but do not fit into one of the above categories. Therefore, in practice *CE-Symm* can also be a good method for identifying, classifying and characterizing solenoid and other repeat proteins with open symmetry. We hypothesize that the low prevalence of actual superhelical symmetry in proteins could be a consequence of the benefit in conserving interfaces between adjacent repeats.

Conclusion

We have extended our internal symmetry analysis tool in order to improve its usability, capabilities and the interpretability of results. In addition to detecting symmetry in protein structures, the tool can identify corresponding residues of the protein from each repeating element and the symmetry operations between them. *CE-Symm* 2.0 adds broad capabilities for the detection of all types of internal symmetry, providing information about the type and order of symmetry and the repeat boundaries. The alignments between the repeats are eminently useful in identifying conserved and differential features between repeats, and can be applied to understanding protein function and evolution.

Determining whether the high prevalence of internal symmetry in protein structures is predominantly a consequence of thermodynamic selection or an indication of the history of protein evolution remains an open question. Here, we have presented examples where internal symmetry is a result of evolution and tied to functional consequences, and how our tool can help researchers in the protein evolution, classification and annotation fields. We have made *CE-Symm* 2.0 and its source code freely available on GitHub, and we are working to integrate it into leading bioinformatics resources for protein analysis.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

P.B. initiated the CE-Symm project and provided direction. S.B., A.L, P.R. and A.P. developed code for the *CE-Symm* tool. S.B. and A.L. designed, implemented and benchmarked the algorithmic extensions of the method and performed the analyses presented in this article. S.B., A.L. and G.C. wrote the manuscript. All authors reviewed and approved the final version of the manuscript.

Acknowledgements

We dedicate this article to Guido, catalyst and indispensable part of this project, who sadly left us before its completion.

We thank Philippe Youkharibache for testing and suggesting new features for *CE-Symm* and Jose Duarte for useful discussions during the development of the method. We also thank the developers that contributed to the *BioJava* library, which has been fundamental throughout the development of *CE-Symm*.

This research was supported in part by the Intramural Research Program of the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (support to S.B and P.B), and by the National Science Foundation, National Institutes of Health, and U.S. Department of Energy [DBI-1338415] (support to A.P., P.R., and P.B.). Financial support to G.C. from the Swiss National Science Foundation and the Research Committee of the Paul Scherrer Institute is gratefully acknowledged. Additional support for S.B from COST Action BM1405 and COST Switzerland SEFRI project IZCNZ0-174836.

References

Abraham, A. L., J. Pothier, and E. P. C. Rocha
2009. Alternative to Homo-oligomerisation: The Creation of Local Symmetry in Proteins by Internal Amplification. *Journal of Molecular Biology*, 394(3):522–534.

Abraham, A. L., E. P. C. Rocha, and J. Pothier
2008. Swelfe: A detector of internal repeats in sequences and structures. *Bioinformatics*, 24(13):1536–1537.

Andrade, M. A., C. Perez-Iratxeta, and C. P. Ponting
2001. Protein repeats: structures, functions, and evolution. *Journal of Structural Biology*, 134(2-3):117–131.

Aravind, P., A. Mishra, S. K. Suman, M. K. Jobby, R. Sankaranarayanan, and Y. Sharma
2009. The gamma-crystallin superfamily contains a universal motif for binding calcium. *Biochemistry*, 48(51):12180–12190.

Bliven, S. E.
2015. *Structure-Preserving Rearrangements: Algorithms for Structural Comparison and Protein Analysis*. PhD thesis, University of California San Diego.

Bliven, S. E., P. E. Bourne, and A. Prlić
2015. Detection of circular permutations within protein structures using CE-CP. *Bioinformatics*, 31(8):1316–1318.

Bliven, S. E., A. Lafita, A. Parker, G. Capitani, and J. M. Duarte
2017. Automated evaluation of quaternary structures from protein crystals. *bioRxiv*, P. 224717.

Bonjack-Shterengartz, M. and D. Avnir
2015. The near-symmetry of proteins. *Proteins: Structure, Function and Bioinformatics*, 83(4):722–734.

Broom, A., A. C. Doxey, Y. D. Lobsanov, L. G. Berthin, D. R. Rose, P. L. Howell, B. J. McConkey, and E. M. Meiering
2012. Modular evolution and the origins of symmetry: Reconstruction of a three-fold symmetric globular protein. *Structure*, 20(1):161–171.

Capitani, G., J. M. Duarte, K. Baskaran, S. Bliven, and J. C. Somody
2015. Understanding the fabric of protein crystals: Computational classification of biological interfaces and crystal contacts. *Bioinformatics*, 32(4):481–489.

Do Viet, P., D. B. Roche, and A. V. Kajava
2015. TAPO: A combined method for the identification of tandem repeats in protein structures. *FEBS Letters*, 589(19):2611–2619.

Goodsell, D. S. and A. J. Olson
2000. Structural Symmetry and Protein Function. *Annu. Rev. Biophys. Biomol. Struct*, 29(1):105–53.

Guda, C., E. Scheeff, P. E. Bourne, and I. Shindyalov
2001. A New Algorithm for the Alignment of Multiple Protein Structures Using Monte Caro Optimization. *Pacific Symposium on biocomputing*, 6:275–286.

Guerler, A., C. Wang, and E. W. Knapp
2009. Symmetric structures in the universe of protein folds. *Journal of Chemical Information and Modeling*, 49(9):2147–2151.

Han, J. H., S. Batey, A. A. Nickson, S. A. Teichmann, and J. Clarke
2007. The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology*, 8(4):319–330.

Jacob, F.
1977. Evolution and tinkering. *Science*, 196(4295):1161–1166.

Karney, C. F. F.
2007. Quaternions in molecular modeling. *Journal of Molecular Graphics and Modelling*, 25(5):595–604.

Kelman, Z. and M. O'Donnell
1995. Structural and functional similarities of prokaryotic and eukaryotic DNA polymerase sliding clamps. *Nucleic Acids Res*, 23(18):3613–3620.

Kim, C., J. Basner, and B. Lee
2010. Detecting internally symmetric protein structures. *BMC bioinformatics*, 11:303.

Kinoshita, K., a. Kidera, and N. Go
1999. Diversity of functions of proteins with internal symmetry in spatial arrangement of secondary structural elements. *Protein science : a publication of the Protein Society*, 8:1210–1217.

Kobe, B. and A. V. Kajava
2000. When protein folding is simplified to protein coiling: The continuum of solenoid protein structures. *Trends in Biochemical Sciences*, 25(10):509–515.

Lee, J. and M. Blaber
2011. Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proceedings of the National Academy of Sciences of the United States of America*, 108:126–130.

Leipe, D. D., L. Aravind, and E. V. Koonin
1999. Did DNA replication evolve twice independently? *Nucleic Acids Res*, 27(17):3389–3401.

Levy, E. D., J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann
2006. 3D complex: A structural classification of protein complexes. *PLoS Computational Biology*, 2(11):1395–1406.

Levy, E. D. and S. Teichmann
2013. Structural, evolutionary, and assembly principles of protein oligomerization. *Progress in Molecular Biology and Translational Science*, 117:25–51.

Lupas, A. N., C. P. Ponting, and R. B. Russell
2001. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of structural biology*, 134(2-3):191–203.

Marsh, J. a., H. a. Rees, S. E. Ahnert, and S. a. Teichmann
2015. Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nature communications*, 6:6394.

Murray, K. B., W. R. Taylor, and J. M. Thornton
2004. Toward the detection and validation of repeats in protein structure. *Proteins: Structure, Function and Genetics*, 57(2):365–380.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia
1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540.

Myers-Turnbull, D., S. E. Bliven, P. W. Rose, Z. K. Aziz, P. Youkharibache, P. E. Bourne, and A. Prlić
2014. Systematic detection of internal symmetry in proteins using CE-symm. *Journal of Molecular Biology*, 426(11):2255–2268.

Pidugu, L. S., K. Maity, K. Ramaswamy, N. Surolia, and K. Suguna
2009. Analysis of proteins with the 'hot dog' fold: Prediction of function and identification of catalytic residues of hypothetical proteins. *BMC Structural Biology*, 9.

Qin, Y.-M., A. M. Haapalainen, S. H. Kilpeläinen, M. S. Marttila, M. K. Koski, T. Glumoff, D. K. Novikov, and J. K. Hiltunen
2000. Human peroxisomal multifunctional enzyme type 2. Site-directed mutagenesis studies show the importance of two protic residues for 2-enoyl-CoA hydratase 2 activity. *Journal of Biological Chemistry*, 275(7):4965–4972.

Shih, E. S. C., R. C. R. Gan, and M. J. Hwang
2006. OPAAS: A web server for optimal, permuted, and other alternative alignments of protein structures. *Nucleic Acids Research*, 34(WEB. SERV. ISS.).

Shindyalov, I. N. and P. E. Bourne
1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering Design and Selection*, 11(9):739–747.

Wolynes, P. G.
1996. Symmetry and the energy landscapes ofbiomolecules. *Proceedings of the National Academy of Sciences of the United States of America*, 93(25):14249.

Zhang, Y. and J. Skolnick
2004. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics*, 57(4):702–710.