Open access • Journal Article • DOI:10.1038/NMETH.2644

Exploiting tertiary structure through local folds for crystallographic phasing

Massimo Sammito, Claudia Millán, Dayté D Rodríguez, Iñaki M. de Ilarduya ...+8 more authors

Institutions: University of Salamanca, University of the Basque Country, University of Göttingen, Catalan Institution for Research and Advanced Studies

Published on: 01 Nov 2013 - Nature Methods (Nature Publishing Group)

Topics: Protein tertiary structure and Protein Data Bank

NATURE METHODS | **BRIEF COMMUNICATION**

Exploiting tertiary structure through local folds for crystallographic phasing

Massimo Sammito,Claudia Millán,Dayté D Rodríguez,Iñaki M de Ilarduya,Kathrin Meindl,Ivan De Marino,Giovanna Petrillo,Rubén M Buey,José M de Pereda,Kornelius Zeth,George M Sheldrick& Isabel Usón

Nature Methods **10**, 1099–1101 (2013) doi:10.1038/nmeth.2644
Received 19 October 2012 Accepted 22 July 2013 Published online 15 September 2013

We describe an algorithm for phasing protein crystal X-ray diffraction data that identifies, retrieves, refines and exploits general tertiary structural information from small fragments available in the Protein Data Bank. The algorithm successfully phased, through unspecific molecular replacement combined with density modification, all-helical, mixed alpha-beta, and all-beta protein structures. The method is available as a software implementation: Borges.

Subject terms: Data processing Protein analysis Proteins X-ray crystallography

Main

With structural knowledge available of the over 80,000 macromolecular crystal structures recorded in the Protein Data Bank (PDB)[1], it should be feasible to solve the 'crystallographic phase problem' for any new structure through computation[2]. The crystallographic phase problem arises because only the diffracted intensities and not the phases are determined from the X-ray diffraction experiment, but the missing phases are essential to compute the structure. Initial phases are usually derived from measurement of heavy atom or anomalous scatterer derivatives, which involves an increase in the experimental effort and timescale of the crystallographic study, as many derivatives turn out to be unsuccessful. Molecular replacement phasing[3, 4], on the other hand, works by locating a related model within the crystallographic unit cell to best account for the experimental diffraction data. Typically, homologs for molecular replacement are retrieved by finding closely related sequences. More recently, molecular replacement using remote homologs has been made successful by combining modeling with the program Rosetta[5]. This requires composing a fairly complete structural hypothesis within a 1.5-Å r.m.s. deviation from the true structure. Alternatively, as little as 10% of the total main-chain structure is enough to achieve phasing at 2-Å resolution, provided that it is almost identical to part of the target structure and accurately placed (r.m.s. deviation <0.5 Å)[6]. Our previous program ARCIMBOLDO[6], for *ab initio* phasing from the native data alone, combines fragment location with Phaser[7] and density modification and autotracing with SHELXE[8] in a supercomputing environment[9]. By applying secondary-structure constraints and density modification[10], it overcomes the resolution and size limitations of direct methods based on constraints derived from atomicity[11]. By sequentially searching for polyalanine helices, ARCIMBOLDO generates hypotheses without specific previous structural knowledge that, if close enough to the true structure, can be expanded to a full solution. The limiting condition is that the search for the first fragment must contain a correct solution; this becomes increasingly challenging for larger structures as the signal becomes weaker.

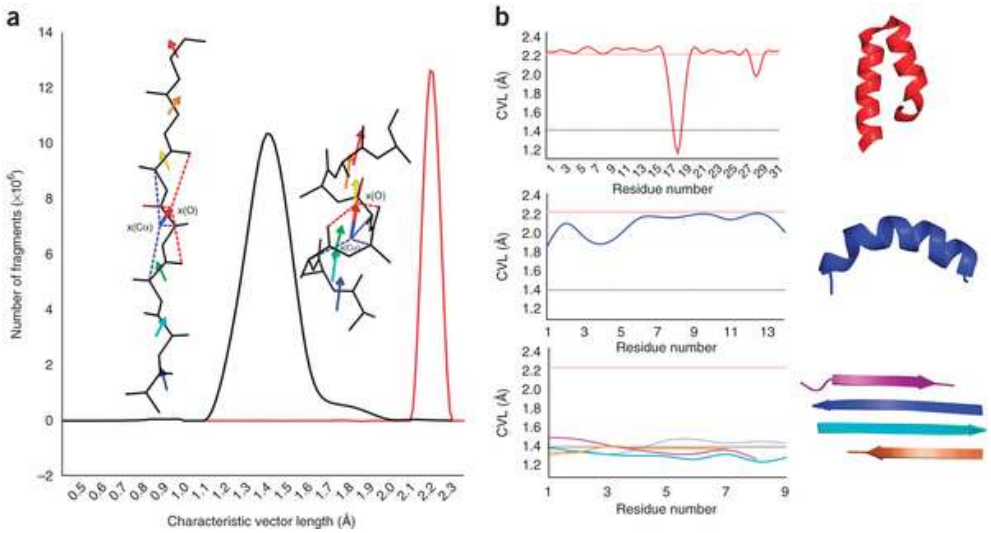
One possible solution is to locate larger, composite fragments, by using tertiary rather than secondary structure. In this particular scenario, modeling has limited use, as both the sequence and the context of the fragment are unknown and optimization would be largely underdetermined. We describe an algorithm and software tool, Borges (http://chango.ibmb.csic.es/BORGES/), that uses tertiary-structure searching in the PDB to solve the crystallographic phase problem.

The PDB contains a vast amount of information, and for any unknown structure, given small enough fragments (for example, two helices or three strands in a particular configuration), close geometrical models are bound to occur in some of the deposited entries. In analogy to Jorge Luis Borges' “Library of Babel” that enclosed books with all random combinations of letters and therefore held any possible book, we reasoned that the information to solve the phase problem is already present in the PDB. All the more so, as unlike the Library of Babel, the PDB is nonrandom, containing in all sorts of structural contexts only the structural units that are stable enough to exist. Further, our method requires small 'sentences' instead of 'volumes', that is, a small fraction of perfect main-chain rather than a complete description of the structure.

Borges runs on a workstation that automatically accesses and distributes calculations to a cluster or supercomputer (Online Methods). Existing tools to analyze and retrieve structural information[12] are meant to identify overall, rather than local, geometry or focus on libraries to be exploited in conventional molecular replacement[13], model building and map interpretation[14] and refinement[15].

In contrast, our approach is tailored to analyze detailed secondary-structure geometry through a distribution of vectors defined by the centroids of alpha carbons and carbonyl oxygens from overlapping tripeptides (Fig. 1a). This distribution accurately discriminates local characteristics, such as sharp kinks or smooth curvature in helices or strands (Fig. 1b). The scalar product of the vectors characterizes relative orientation and is used to compare and cluster results, retrieving from the PDB a comprehensive library of composite main-chain fragments within a tolerance of our geometrical definition and independently of the sequence. Our algorithm is devised to characterize the structural landscape of local folds. Borges identifies, retrieves, clusters, refines and exploits this geometric information, guided by the experimental diffraction data to solve structures where single-fragment search fails.

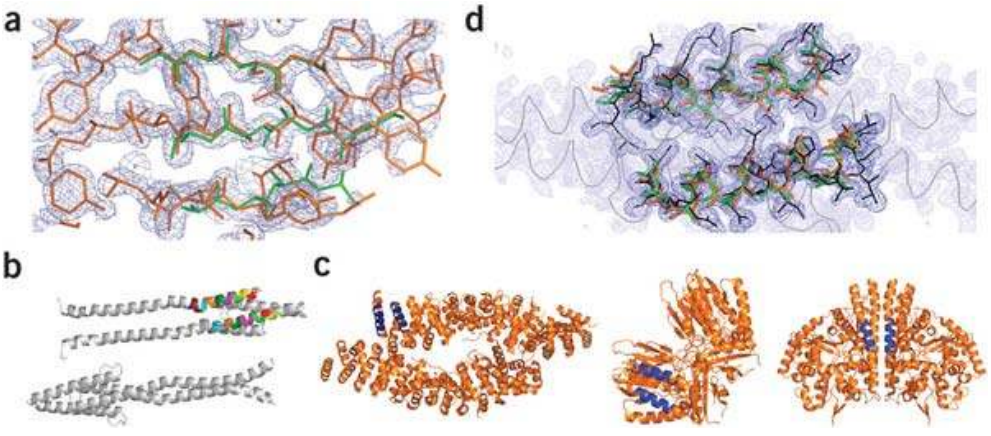
Figure 1: Characteristic Cα-O vectors (CVs) used in Borges to handle secondary structure and local fold geometry.



(**a**) Occurrence of CV values for all β-strands (black) and α-helical (red) tripeptides in the PDB. CVs are shown for each amino acid on a β-strand (left structure) and an α-helix (right structure); the central one highlights the Cα centroid (X(Cα), blue) and O centroid (X(O), red) defining the vector. (**b**) CV length (CVL) distribution for the fragments displayed. Top, helical structure disrupted by a sharp and a light kink. Center, curved helix showing CVs below the red line representing perfect helicity, consistent with local distortions. Bottom, CVL distribution for the four strands in the β-sheet displayed with colors matching the cartoon plot of the structure at right. Again, distortion brought up by sheet curvature causes a decrease in CV, with the black line representing average CV value for β-strands. Figures were prepared with Gnuplot (http://www.gnuplot.info/) and PyMOL (http://www.pymol.org/).

We applied our method to a number of test cases (Supplementary Results) comprising all-helical proteins, mixed alpha-beta proteins and an all-beta protein (rei7, PDB 4L1H, an immunoglobulin domain) (Fig. 2a). Supplementary Table 1 reports the search fragments used to phase these structures; two parallel or antiparallel α-helices or three-stranded antiparallel, parallel and parallel antiparallel β-sheets of 13–21 residues were found to be suitable search fragments. We also solved the previously unknown partial structure of the membrane protein AF1503 from *Archaeoglobus fulgidus*[16] (PDB 4GN0), a three-domain protein monomer composed of an N-terminal extracellular GAF-like domain, a membrane-spanning helix and a C-terminal Hamp domain. The crystals, containing dimers of the Hamp domain and an adjacent helical part (411 amino acids), diffracted to 1.75 Å (Supplementary Table 2).

Figure 2: Overall occurrence of model fragments and their role in phasing an all-beta and a previously unknown structure.



(**a**) Solving fragment (green) and SHELXE electron density map (*F*-weighted mean phase error = 20°) for all-beta 4L1H. (**b**) Cartoon representation of 4GN0; colored regions display the location of fragments leading to solution. (**c**) Three of the 13 structures from which

models (blue) were extracted to solve the 4GN0 structure: β-catenin–BCL9–Tcf4 complex (2GL7), bacterial Mre11 core (3THN) and cytochrome C nitrite reductase (1QDB). (**d**) SHELXE electron density map (FwMPE = 42°). The final 4GN0 model is depicted in gray. The closest original fragment (orange) extracted from 2GL7 (ref. 19) has an r.m.s. deviation of 0.90 Å, whereas after refinement (green) the r.m.s. deviation is 0.54 Å.

For AF1503, we searched for sets of two contiguous parallel α-helical polyalanine fragments of 16 residues, of similar geometry (within 5 Å r.m.s. deviation) to a model cut from the soluble domain of a membrane-anchored protein, and used the resulting clustered library, which contained 460,000 models (Supplementary Results). All successful fragments targeted the same area of the structure (Fig. 2b). Phasing was achieved with 14 models derived from unrelated structures (Fig. 2c). None of these fragments would have succeeded in phasing without our algorithm interspersing geometrical refinement against the diffraction data: this allowed the model to change after rotation search and before density modification. Refinement in the triclinic cell is essential for the translation function to succeed. Refinement against the rotation function[17, 18] has been previously used within molecular replacement, Patterson correlation refinement being part of the molecular replacement–Rosetta approach[5]. Its performance may be irregular, but as the method cycles and selects promising solutions according to their figures of merit, improved models that succeed in the translation are prioritized for further stages. Models are refined again after the translation search to allow small enough deviations to the true structure for density modification and autotracing to succeed. Likewise, model trimming to optimize the correlation coefficient is essential in the case of β-strands. Indeed, even for such a reduced model as two parallel helices the target local fold that needs to be found was unexpectedly unique in the case of AF1503. The closest fragment in the clustered library is too different from its target to succeed (Fig. 2d). The whole PDB contains only six fragments under 0.6 Å r.m.s. deviation, but all of these fragments are from structures that are related to the target by sequence.

Not all possible fragments of two parallel helices extracted from the final structure can be located and succeed in phasing (Supplementary Fig. 1), which is hardly surprising because contribution to diffraction is not uniform throughout a structure, with more rigid parts being more effective. Model refinement against the experimental data and selection through figures of merit (Supplementary Results) drives calculations toward the most prominent features that allow phasing. The implementation of statistical knowledge about which building blocks are both more common and most rigid into Borges could further narrow the search.

For the unknown 223-amino-acid structure of a plectin fragment 4GDO at 1.7 Å in *C*2 (Supplementary Table 2 and Supplementary Fig. 2), the structural fragment needed for phasing was already found in the PDB. 967 models out of 121 structures (82 unrelated) in the PDB were similar enough to one portion of the final structure (<0.6 Å) for our library of contiguous antiparallel helices to solve, even without further model refinement.

In conclusion, we solved both test-case and unknown structures displaying diverse folds by using tertiary-structure constraints for phasing. Our method, implemented in Borges, exploits previously described structural building blocks by extracting comprehensive collections of small composite main-chain fragments from the PDB regardless of sequence. Even for such a minimal unit as two helices, unknown structures may contain either frequent or unique local folds so that a pure brute-force phasing method would not be successful, but we show that the experimental data can drive selection and refinement of the composite model fragments toward the target structure.

Methods

Software availability.

Borges is available at http://chango.ibmb.csic.es/BORGES/.

Along with the code, documentation for its installation, setup (Supplementary Note 1) and use (Supplementary Note 2) as well as a tutorial can be downloaded.

Borges was designed and developed to run on a local machine accessing a local or remote Condor/SGE grid environment. The core part of the program will run on a local machine independently of the grid environment available. Access to the remote grid (user, access key, paths and addresses) is input into the configuration files.

The Borges algorithm.

Our algorithm for describing secondary structure and computing tertiary structure relies on a distribution of characteristic vectors (CVs) defined by the centroids of α-carbons and carbonyl oxygens of consecutive, overlapping tripeptides. The backbone conformation of a tripeptide captures an amino acid in the context given by its preceding and its following residue. For every tripeptide along the protein backbone, a vector is defined with its origin at the geometric centroid of the three α-carbon atoms and

ending at the centroid described by the three carbonyl oxygens in the tripeptide. In the case of an α-helix, the CVs are parallel to its axis, and their direction is that of the polypeptide chain; for a β-strand, the CVs deviate around 45° from the direction of the polypeptide chain, with consecutive vectors being approximately orthogonal (Fig. 1a). The moduli of such vectors are determined by the kind of secondary structure, their distributions falling into clearly distinct ranges: the resulting mean (standard deviation) values are alpha, 2.19 (0.18) Å; beta, 1.39 (0.21) Å; or coil (where individual tripeptides may show any CV value, but the distribution along consecutive segments varies erratically). These values reflect the hydrogen bonding undergone by the carbonyl oxygens, and thus the main-chain environment and interactions. The CV modulus is maximized when all carbonyl moieties are aligned, as in the more regular α-helices; distortions from helix bending or kinks are concomitant to a change in the carbonyl orientation, leading to a sensible decrease in the resulting CV modulus. The alternating geometry adopted by the directions of the carbonyl groups in a strand leads to a substantial shortening of the resulting CV. Loop and coil regions tend to contain backbone torsions in the preferred Ramachandran regions, and thus it is not surprising that CVs for their tripeptides may adopt any value but consecutive CVs lack the constant distribution identifying the secondary-structure elements. Beyond secondary structure, CVs are useful to ascertain tertiary structure relationships. To this end, a global CV is defined for the complete N-peptide in each secondary-structure fragment, that is, a vector defined from the centroid of all α-carbons to the centroid of all carbonyl oxygens in the fragment. Distances between different secondary-structure elements can be calculated from these global CVs, through the geometrical difference of their origin points; and through their scalar product their relative orientation can be quantified. This formulation is conveniently accurate, matches DSSP assignment and discriminates well among local characteristics. It also provides the flexibility to define different thresholds for the geometry of different areas (for example, a more rigid definition of strands packed within a β-sheet flanked by a more mobile α-helix).

The geometrical definition of a library is conveyed to Borges through a model template in PDB format and a configuration file. The program analyzes the template and its correspondence to the instructions in the configuration file and prompts the user to resolve any ambiguity or contradiction. With this information, all main-chain composite fragments in the PDB with a tertiary structure resembling the template within the specified thresholds will be extracted and superimposed. Threshold values are defined in the configuration file for the distance between fragments, calculated from the distances relating CV origins. Limits are also defined for the deviation of the angles between fragments: the angles between corresponding CVs in the template and fragment cannot differ by more than a given value for the fragment to be accepted. Finally, a limit is set for the degree that template and fragment may differ in the angles between the distance vector and the global vector of the secondary-structure element at its origin, in order to break the correlation among distance tilt and relative translation introduced by the use of a projection. When the geometry tolerance involving two secondary-structure fragments is assessed, the less restrictive limit will be applied. Thus, if for instance a geometry defining relative positions of three β-strands within 4 Å and of an helix within 8 Å were specified in the configuration file, relationships among strands would be limited by the 4 Å threshold, whereas relationships between the helix and each of the strands would have to observe the 8 Å limit.

Borges needs to compute every characteristic vector for each tripeptide in the PDB before screening for a given geometrical definition. But as both operations are independent, the whole PDB (stand 12 January 2012, updated up to 14 February 2013 for the results discussed) is filtered and annotated in terms of characteristic vectors to produce a database (17 GB) from which different fragment libraries may be derived.

Borges processes each .pdb file provided and, in the case of NMR structures, each model contained. Artificial *B* values are adopted. For nonredundant sets, the database or the search can be limited to a single model per NMR structure. Also, NMR models can be completely filtered out if so wished for a given problem.

Generation of the annotated database took 17 h on a four-core workstation. A search against the resulting database to extract and cluster a given geometry (for example, two helices or three β-strands in a particular disposition) takes under half a day in a grid of 100 cores, with more complex motifs or non-exhaustive samplings being considerably faster.

To extract a library from the annotated database, Borges starts by analyzing the template provided. It decomposes the template into secondary-structure fragments, described by their CV distribution, and computes relative geometrical relationships between them. Let us define X_s and W_s as generic elements of secondary structure, α-helix or β-strand that belong to the search model. If X_s has t residues, Borges associates to this fragment a distribution X of $t - 2$ CVs. The same is done for all other secondary-structure elements W_s. Borges also describes each fragment with a global CV defined as the vector between the centroid of all Cα and the centroid of all O atoms.

$$X_S = R_1^x, R_2^x, \ldots, R_t^x;\ X = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n=t-2}$$

$$W_s = R_1^w, R_2^w, \ldots, R_m^w;\ W = \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{q=m-2}$$

$$\mathbf{W}_{\mathrm{CV}} = \frac{1}{m}\sum_{i=1}^{m} C\alpha(R_i^w) - \frac{1}{m}\sum_{i=1}^{m} \mathrm{O}(R_i^w)$$

Internal relationships between secondary-structure elements are computed: the relative orientation of the two fragments $\mathbf{X}_{\mathrm{CV}}$ and $\mathbf{W}_{\mathrm{CV}}$ is expressed as the angle γ, the distance between them is expressed as a vector connecting the Cα centroids of the two fragments *X* and *W*. The resulting distance vector is named $\mathbf{D}_{xw}$ of length *r*, and the direction of this distance is determined by the angle ϕ between $\mathbf{X}_{\mathrm{CV}}$ and $\mathbf{D}_{xw}$.

$$\gamma = \measuredangle \mathbf{X}_{\mathrm{CV}}, \mathbf{W}_{\mathrm{CV}}$$

$$\mathbf{D}_{xw} = \frac{1}{t}\sum_{i=1}^{t} C\alpha(R_i^x) - \frac{1}{m}\sum_{i=1}^{m} C\alpha(R_i^w)$$

$$\phi = \measuredangle \mathbf{X}_{\mathrm{CV}}, \mathbf{D}_{xw}$$

Once Borges has analyzed all secondary-structure elements in the search model, it creates a graph in which each node is a secondary-structure element that is connected to its nearest node. A robust topological sorting is used, initially identifying the strongly connected components and then performing the topological sorting on these components. This kind of ordering is limited to acyclic graphs, and strongly connected components are acyclic by definition. The ordering procedure allows Borges to identify a core of the fold, represented by the elements that are most closely packed together in space. It is convenient to start by searching in the annotated database those core secondary-structure elements to discard early on incompatible combinations, thereby freeing memory and reducing computation time.

Borges sequentially finds the secondary-structure elements; thus, initially it can apply filters and constraints related to the internal geometry and properties of only the first fragment; after having fixed the first element and searching for a second, it can also apply geometrical filters between fragments.

The geometric conditions are evaluated by checking the less probable conditions first in order to filter out fast geometries that will not fit the template definition.

Filtering criteria. Given Y_s and Z_s, two secondary-structure elements in a structure of the PDB that correspond to secondary-structure elements X_s and W_s defined for the template, the basic descriptions are computed in the same way, defining *Y*, *Z* as the distribution of CVs and $\mathbf{Y}_{\mathrm{CV}}$ and $\mathbf{Z}_{\mathrm{CV}}$ as the global CVs.

Secondary-structure fragment *Y* will be compared to *X* and *Z* to *W*, and the geometrical relationships of the pair *Y*, *Z* will be compared with those in the search model *X*, *W*.

Fragment length. *X* and *Y* must have the same number of residues. So as a consequence they also will have the same number of CVs in their distributions.

$$|X_s| = |Y_s| \Leftrightarrow |X| = |Y|$$

Secondary structure. *X* and *Y* should have the same secondary-structure annotation, verified through the CV distribution. A pair of two successive CVs belonging to the same fragment may not differ in their length by more than 1.0 Å, or else a breakpoint is defined in the secondary-structure element. Then each CV of the distribution is compared with statistical values to check consistency and continuity for ah, α-helices, and bs, β-strands. Whenever outliers are found, before the entire fragment is rejected, a Ramachandran validation is performed for the torsions in the tripeptide, and it is also checked that the predecessor and successor of the outlier CV still belong to the predicted secondary-structure annotation. In that case, a distortion of the fragment is registered, flagging as ch, curved helices, and cbs, strongly distorted strands.

$$\mathrm{type}(X,\mu,\sigma)$$
$$= \begin{cases} \text{if } \forall i \in \{1,\ldots,n\} \\ \qquad |\,|\,\mathbf{x}_i\,|\,| - \mu\,| \le \sigma \Rightarrow \text{regular} \\ \text{if } \exists C \subset \{2,\ldots,n-1\} : \forall j \in C \\ \quad |\,|\,\mathbf{x}_{j-1}\,|\,| - \mu\,| \le \sigma \wedge |\,|\,\mathbf{x}_{j+1}\,|\,| - \mu\,| \le \sigma \Rightarrow \text{distorted} \end{cases}$$
$$\mathrm{type}(X,\mu=2.19,\sigma=0.18) = \text{regular} \Rightarrow \text{ah}$$
$$\mathrm{type}(X,\mu=1.39,\sigma=0.21) = \text{regular} \Rightarrow \text{bs}$$
$$\mathrm{type}(X,\mu=2.19,\sigma=0.18) = \text{distorted} \Rightarrow \text{ch}$$
$$\mathrm{type}(X,\mu=1.39,\sigma=0.21) = \text{distorted} \Rightarrow \text{cbs}$$
$$\mathrm{type}(X,\mu,\sigma) = \mathrm{type}(Y,\mu,\sigma)$$

Scalar product for relative orientation. To check relative orientations between fragments, Borges computes the scalar product of their global CVs given *γ*, the angle between *X* and *W*, and *α*, the threshold on the difference in the corresponding template and fragment angles, specified by the user in the configuration file.

$$\theta = \cos^{-1}\left(\frac{\mathbf{Y}_{\mathrm{CV}} \cdot \mathbf{Z}_{\mathrm{CV}}}{|\,|\,\mathbf{Y}_{\mathrm{CV}}\,|\,|\,|\,\mathbf{Z}_{\mathrm{CV}}\,|\,|}\right)$$

$$\gamma - \alpha \le \theta \le \gamma + \alpha$$

Fragment distance. For checking distance compatibility for *Y* and *Z*, a distance vector $\mathbf{D}_{yz}$ is defined. In the template, *r* is the length of the distance vector $\mathbf{D}_{xw}$, and ϕ is the angle between $\mathbf{X}_{\mathrm{CV}}$ and $\mathbf{D}_{xw}$. Distance and angle have to agree with those in the template within the user-specified thresholds input in the configuration file, named *d* and *β*, respectively

$$r - d \le |\,|\,\mathbf{D}_{yz}\,|\,| = \sqrt{\frac{1}{t}\sum_{i=1}^{t} \mathrm{C}\alpha(R_i^y) - \frac{1}{m}\sum_{i=1}^{m} \mathrm{C}\alpha(R_i^z)} \le r + d$$
$$\phi - \beta \le \measuredangle \mathbf{Y}_{\mathrm{CV}}, \mathbf{D}_{yz} \le \phi + \beta$$

Distribution difference. This filter limits the maximum difference between each CV of the search fragment element *X* and its corresponding CV of the extracted secondary-structure element *Y*, thus enabling the user to define the tolerance threshold, *δ*, within which local geometrical distortion of the fragments may differ. Even if the program will accept any positive number, physically meaningful values range between 0.15 and 0.40 Å; larger values would be comparing different types of secondary-structure elements, previously filtered out. Analyses were performed on all crystallographic structures from the PDB, computing the CV distribution of their overlapping tripeptides. The CV length was correlated with the corresponding DSSP prediction for those residues. Two distinct distributions describe secondary-structure elements, and the range of values falling under either distribution (the amplitudes of the interval are 0.5 and 0.2 Å for β-strands and α-helices, respectively) gives a physical interpretation of the effect that limiting the distribution difference may have in practice.

$$\forall i \in \{1,\ldots,n\} \;\; |\,|\,|\mathbf{x}_i|\,| - |\,|\mathbf{y}_i|\,|\,| \le \delta$$

All models finally extracted from the PDB are grouped in geometrically similar clusters. The r.m.s. deviation among models in a cluster can be chosen by the user; for our phasing purposes, values below 1 Å were found to be effective. The clustering algorithm, based on the enclosure algorithm over the connected-component graph[20], in which each node presents a model, is applied on the fly. The fragment with the lowest r.m.s. deviation from the template is chosen to represent a cluster, but given the low clustering threshold, they all should be geometrically very similar. All fragments representing clusters are superimposed on the template in order to evaluate globally the fragment location results common to subsets of model clusters, and *B* factors are all set to the same value.

Use of the clustered library for phasing.

Each cluster representative becomes a search model that is examined in parallel through fast rotation with Phaser at low resolution (3 Å), as most of the models will present large r.m.s. deviations compared to the target structure. All rotation solutions obtained

(typically 200) for each of the models (thousands) are analyzed so that they can be grouped in clustered rotations common to a subset of models. They are ranked after their rotation figures of merit and population of models to further pursue them sequentially and independently. First, a brute-force rotation around the clustered rotation is performed. This is followed by rigid-body refinement of individual fragments against the rotation function, fast translation, brute-force translation to improve accuracy of the location, packing filtering and rigid body refinement. Finally, the distribution of initial fragment correlation coefficients after fragment trimming is analyzed to select the partial solutions sent to iterative density modification and autotracing. The outlined procedure is dynamic and adapts to the crystallographic particularities of the case, such as skipping translation searches if the space group is *P*1 or packing checks in space groups without screw axes or Patterson correlation refinement, depending on the models or use of common rotation results for enantiomeric space group pairs.

Expression, purification, crystallization and data collection of a fragment of the rod domain of plectin.
The cDNA sequence coding for residues 1382–1420 of human plectin (UniProtKB accession number Q15149-2) was amplified by PCR and was cloned into the pGEX-4T3 vector (GE Healthcare). The plectin fragment was expressed as a glutathione S-transferase (GST) fusion protein in *Escherichia coli* strain BL21(DE3)T1 and was purified using glutathione Sepharose (GE Healthcare) affinity chromatography followed by overnight in-column digestion with thrombin at room temperature. The cleaved plectin protein was dialyzed against the desired buffer and concentrated using Amicon ultrafiltration cells (Millipore). Crystals were obtained by sitting-drop vapor diffusion at room temperature by mixing a protein solution at 25 mg/ml in 5 mM Tris-HCl pH 7.5 with an equal volume of the crystallization solution 0.1 M sodium acetate (pH 4.5), 0.2 M Li_2SO_4 and 2.3 M NaCl. Prior to data collection, a crystal was transferred into Paratone-N oil and was cooled by immersion in liquid nitrogen. Data were collected at 100 K using a Microstar-H rotating anode (Bruker AXS) and a mar345 detector (Marresearch GmbH). Diffraction intensities were integrated, reduced and converted into structure factor amplitudes with the XDS suite[21].

Expression, purification and crystallization of a fragment of the membrane protein AF1503CC.
The modified protein AF1503 from *A. fulgidus* was cloned into a pet30b vector and was expressed in BL21 (DE3) gold cells. Expression was performed at 37 °C and induction at an OD of 0.6 using 1 mM IPTG. The protein was purified by anion-exchange chromatography (QHP, 21 ml; GE Healthcare) in 30 mM MOPS, pH 7, via a linear gradient of 50–525 mM NaCl. The eluted protein was precipitated using 30% ammonium sulfate on ice and subsequently resuspended in 20 mM MOPS, 100 mM NaCl, 10 mM EDTA, pH 7. AF1503CC was further purified via size-exclusion chromatography using a Superdex column (Superdex S200, 26/60; GE Healthcare) and 20 mM MOPS and 100 mM NaCl as running buffer. The fractions containing AF1503 were collected and concentrated to 10 mg/ml for crystallization. Crystallization of the protein was performed using 800 conditions of commercial screens (Hampton Research, Qiagen). Crystals appeared under a variety of conditions, but only a minor fraction of the crystals tested diffracted to high resolution. The best crystals were obtained in 25% PEG3350 and 100 mM HEPES, pH 6.5. Crystals were frozen either directly or after addition of 10% PEG400 to the reservoir solution. Data were collected at the PX10 beamline of the SLS (Swiss Light Source), Villigen, Switzerland. Data were recorded at 100 K on a Pilatus detector 6M (Dectris) at 20% intensity of the full beam (400 mA), and data were processed using the XDS/XScale program package[21].

Accession codes.
PDB: 4GN0, 4GDO, 4L1H.

Accession codes

Primary accessions

Protein Data Bank

4L1H	4L1H
4GN0	4GN0
4GDO	4GDO

Referenced accessions

Protein Data Bank		Swiss-Prot
2GL7	2GL7	Q15149-2
3THN	3THN	
1QDB	1QDB	

References

1. Bernstein, F.C. *et al*. *J. Mol. Biol.* **112**, 535–542 (1977).

2. Qian, B. *et al*. *Nature* **450**, 259–264 (2007).
3. Huber, R. *Acta Crystallogr.* **19**, 353–356 (1965).
4. Rossmann, M.G. *The Molecular Replacement Method* (Gordon and Breach, 1972).
5. DiMaio, F. *et al*. *Nature* **473**, 540–543 (2011).
6. Rodríguez, D.D. *et al*. *Nat. Methods* **6**, 651–653 (2009).
7. McCoy, A.J. *et al*. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
8. Sheldrick, G.M. *Acta Crystallogr.* **D66**, 479–485 (2010).
9. Tannenbaum, T., Wright, D., Miller, K. & Livny, M. in *Beowulf Cluster Computing with Linux* (ed. Sterling, T.) Ch. 14, 307–350 (The MIT Press, 2002).
10. Burla, M.C., Carrozzini, B., Cascarano, G.L., Giacovazzo, C. & Polidori, G. *J. Appl. Crystallogr.* **44**, 1143–1151 (2011).
11. Miller, R. *et al*. *Science* **259**, 1430–1433 (1993).
12. Joosten, R.P. *et al*. *Nucleic Acids Res.* **39**, D411–D419 (2011).
13. Oldfield, T.J. *Acta Crystallogr.* **D57**, 1421–1427 (2001).
14. Cowtan, K. *Acta Crystallogr.* **D68**, 328–335 (2012).
15. Nicholls, R.A., Long, F. & Murshudov, G.N. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 404–417 (2012).
16. Linder, J.U. & Schultz, J.E. *Methods Enzymol.* **471**, 115–123 (2010).
17. Brünger, A.T. *Methods Enzymol.* **276**, 558–580 (1997).
18. Grosse-Kunstleve, R.W. & Adams, P.D. *Acta Crystallogr.* **D57**, 1390–1396 (2001).
19. Sampietro, J. *et al*. *Mol. Cell* **24**, 293–300 (2006).
20. Hopcroft, J. & Tarjan, R. *Commun. ACM* **16**, 372–378 (1973).
21. Kabsch, W. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

Download references

Acknowledgments

This work was supported by the Spanish Ministerio de Ciencia e Innovación-Ministerio de Economia y Competitividad, Centro de Desarrollo Tecnológico Industrial and Consejo Superior de Investigaciones Científicas (grants BIO2009-10576; IDC-2010-1173; BFU2012-35367; BFU2012-32847; predoctoral grants to D.D.R., I.D.M. and I.M.d.I.; JdC to K.M.; RyC to R.M.B.); Generalitat de Catalunya (2009SGR-1036); VW-Stiftung Niedersachsenprofessur to G.M.S. We also acknowledge beam time on the Swiss Light Source beamline X10SA and computing time at the FCSCL.

Author information

Affiliations

Instituto de Biología Molecular de Barcelona, Consejo Superior de Investigaciones Científicas, Barcelona, Spain.
Massimo Sammito,Claudia Millán,Dayté D Rodríguez,Iñaki M de Ilarduya,Kathrin Meindl,Ivan De Marino,Giovanna Petrillo &Isabel Usón

Instituto de Biología Molecular y Celular del Cáncer, Consejo Superior de Investigaciones Científicas–Universidad de Salamanca, Salamanca, Spain.
Rubén M Buey &José M de Pereda

Ikerbasque Basque Foundation for Science at Unidad de Biofísica, Consejo Superior de Investigaciones Científicas–Universidad del País Vasco, Leioa, Spain.

Kornelius Zeth

Lehrstuhl für Strukturchemie, Universität Göttingen, Göttingen, Germany.
George M Sheldrick

Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.
Isabel Usón

Contributions
All authors contributed extensively to the work presented in this paper.

Competing financial interests
The authors declare no competing financial interests.

Corresponding author
Correspondence to: Isabel Usón

Supplementary information

PDF files

1. Supplementary Text and Figures (6,597 KB)
   Supplementary Figures 1 and 2, Supplementary Tables 1 and 2, Supplementary Results and Supplementary Notes 1 and 2

Nature Methods ISSN 1548-7091 EISSN 1548-7105