nature | methods

Crystallographic *ab initio* protein structure solution below atomic resolution

Dayte D Rodríguez, Christian Grosse, Sebastian Himmel, César González, Iñaki M de Ilarduya,
Stefan Becker, George M Sheldrick & Isabel Usón

Supplementary figures and text:

Supplementary Table 1	Data collection and refinement statistics (*ab Initio* phasing) for pdb entry 3GWH
Supplementary Results	Test protein results
Supplementary Figure 1	Location of $(Ala)_{12}$ helices for the structure of glucose isomerase.
Supplementary Figure 2	Ab Initio Fo.fom electron density map for EIF5 derived from 4 ideal polyalanine helices of 12 residues after density modification.

Nature Methods: doi:10.1038/nmeth.1365

Supplementary Table 1 Data collection and refinement statistics (*ab Initio* phasing) for pdb entry 3GWH

	PRD-II native data
Data collection	
Space group	$P2_1$
Cell dimensions	
a, *b*, *c* (Å)	37.39, 65.75, 38.19
β (°)	109.58
Resolution (Å)	24.0-1.95 (2.0-1.95)*
R_{sym}	4.64 (41.8)
$I / \sigma I$	18.8 (2.25)
Completeness (%)	99.7 (97.6)
Redundancy	15.5 (2.6)
Refinement	
Resolution (Å)	24.03-1.95 (2.0-1.95)
No. reflections	12832 (893)
R_{work} / R_{free}	19.5/24.1 (31.8/ 25.4)
No. atoms	1680
Protein	1641
Ligand/ion	10
Water	29
B-factors	27
Protein mainchain /sidechain	25 / 28
Ligand/ion	40
Water	30
R.m.s. deviations	
Bond lengths (Å)	0.023
Bond angles (°)	1.848

*Data from one twinned crystal was used. *Values in parentheses are for highest-resolution shell.

Supplementary Results

1.1. CopG: learning from the atomic resolution case

The asymmetric unit of the orthorhombic C222$_1$ CopG contains three monomers of 43 amino acids each (1015 protein atoms in total) and 40% solvent. Six helices of (~12 amino acids) are present. Synchrotron data to a maximum resolution of 1.2 Å are available, but an *ab Initio* solution with SHELXD has not been achieved.

Using default parameters but full resolution for an automatic PHASER run, with a theoretical model helix made up of 10 alanines, combining rotation, translation, packing and refinement yielded 4 correctly placed helices and failed to find the fifth and sixth helices. The reason is that the default search grid is too coarse when dealing with such small fragments. It can be overcome by adopting a mesh of 1-3 degrees for the rotation and 0.5-0.7 Å for the translation. An even finer grid for the translation search requires more memory than our computers can provide. These values are used in all following tests.

Nevertheless, the 4 located fragments are enough start information for SHELXE, to obtain phases that would permit to build the whole structure. Even 2 out of the 20 partial solutions containing a single helix of 10 aminoacids are enough to phase the structure (MPE: 29.1°) but they are not the ones with the highest figures of merit, log-likelihood gain (LLG). Thus, if even at atomic resolution correct solutions cannot be perfectly discriminated by their figures of merit, a multisolution frame is imperative to increase the chances of structure solution. The 4 correctly placed helices can be extended judging on the LLG. Rescoring LLG on a set of model helices of different lengths superimposed on the 10 aminoacids search model, allows to extend the original fragment if the LLG value increases, thus completing three helices of 14 aminoacids and one of 12 (LLG= 98 for the initial 40 residues substructure, LLG= 136 for the improved 54 residues one). After thus completing the model, helices 5 and 6 can be found as well.

The effect of resolution on our tests appears somewhat erratic. One would naturally expect that using the highest possible resolution would lead to optimal results, but this is not the case and our previous work anticipated it[19]. One possible explanation could lie in the fact that the resolution shells around 2.4 Å contain the values of most 1-3 bonded interatomic distances in a macromolecule and their projections perpendicular to the diffraction planes. This is reflected in the local maximum shown in the distribution of average squared normalized structure-factor amplitudes as a function of resolution in this region[20]. In the case of CopG, truncating the resolution of the data from the available 1.2 to 2.1 Å for the rotation function still yields correctly oriented (Ala)10 helices, for which the translation function works well truncating the data to 1.5 Å, whereas no translation solution is achieved for these rotations when data are truncated to either 1.8 Å or 2.1 Å. On the other hand, performing the rotation search with data truncated to 1.8 Å and using these rotations to determine translations with data truncated to 2.1 did work, yielding a complete solution from the expansion of some substructures composed of 2 fragments. Truncating the resolution for both rotation and translation search to 2.5 Å did not yield useful substructures for phasing.

1.2. Glutaredoxin: solution of a small protein at 1.45 Å with more strands than helices

The asymmetric unit of the orthorhombic $P2_12_12_1$ 1ABA PDB entry contains 87 amino acids (880 atoms including solvent and ligands) and 45% solvent. Three helices of (10, 12 and 14 amino acids) are present. Data to a maximum resolution of 1.45 Å are available.

The fragment search was set up to locate 3 helices of 10 alanines, restricting the resolution for the rotation search to 2.1 Å and using data to the full 1.45 Å resolution for the translation search and rigid body refinement with PHASER. The rotation search was carried out in 2º steps and translation in 0.7 Å steps. Combining every rotation peak with every translation peak would soon lead to an unmanageable number of solutions. In order to control the flow of the search, the following limits were set: For every rotation or translation search, peaks under 75% of top were rejected, as is the default in PHASER. Furthermore, from each translation run after the first fragment, no more than 70 solutions were further pursued. After the packing check, surviving substructures were divided in groups of 400 solutions for rigid body refinement and pruning of duplicates. Only the 100 top solutions were kept. Expansion to the full structure with SHELXE works with substructures made up of 1, 2 and 3 helices. The structure was solved starting from 1 helices (50 atoms), in 3.6 % of the cases, starting with 2 helices in 6.6% of the cases and from 3 helices in 15% of the substructures. In every SHELXE attempt, starting from phases derived from the partial structure, 5 runs of density modification made up of 30 cycles each were interspersed with autotracing. The density sharpening parameter (v) was set to 0 and reflections were extrapolated to a resolution of 1 Å.

1.3. GI, Tests on a TIM barrel protein at 1.54 Å resolution

Of particular interest is the location of model fragments when the resolution of the data is high but worse than atomic and the size of the protein with no heavier elements than sulfur present exceeds a thousand atoms, as under such circumstances dual-space recycling *ab Initio* is not expected to work. In this test, data collected in-house for Glucose Isomerase to a maximum resolution of 1.54 Å were used. The asymmetric unit contains 368 amino acids, so both for its size and resolution the problem would be well beyond the reach of current *ab Initio* methods. The protein fold is that of a TIM barrel, so that it contains a large proportion of helical structure, including some long helices (over 20 amino acids). For this structure, a complete helix of 23 aminoacids (186 atoms) perfectly positioned is enough to provide starting phases that will allow SHELXE to solve the structure (the MPE creeps down from 77º to 25º in the course of 1000 cycles). In practice, more than a single helix will be needed, as it will be neither perfectly positioned nor complete (with side chains). But as a start, the location of such a fragment (residues 150 –172 from the final refined structure) was undertaken.

Using the full resolution of the data for the rotation search did not render the correct orientation in a rotation search, whereas this could be found truncating the data to 2.1 Å, albeit not as the first solution (83% of top peak). When the rotation search is undertaken using data up to 2.6 Å the rotation search also fails.

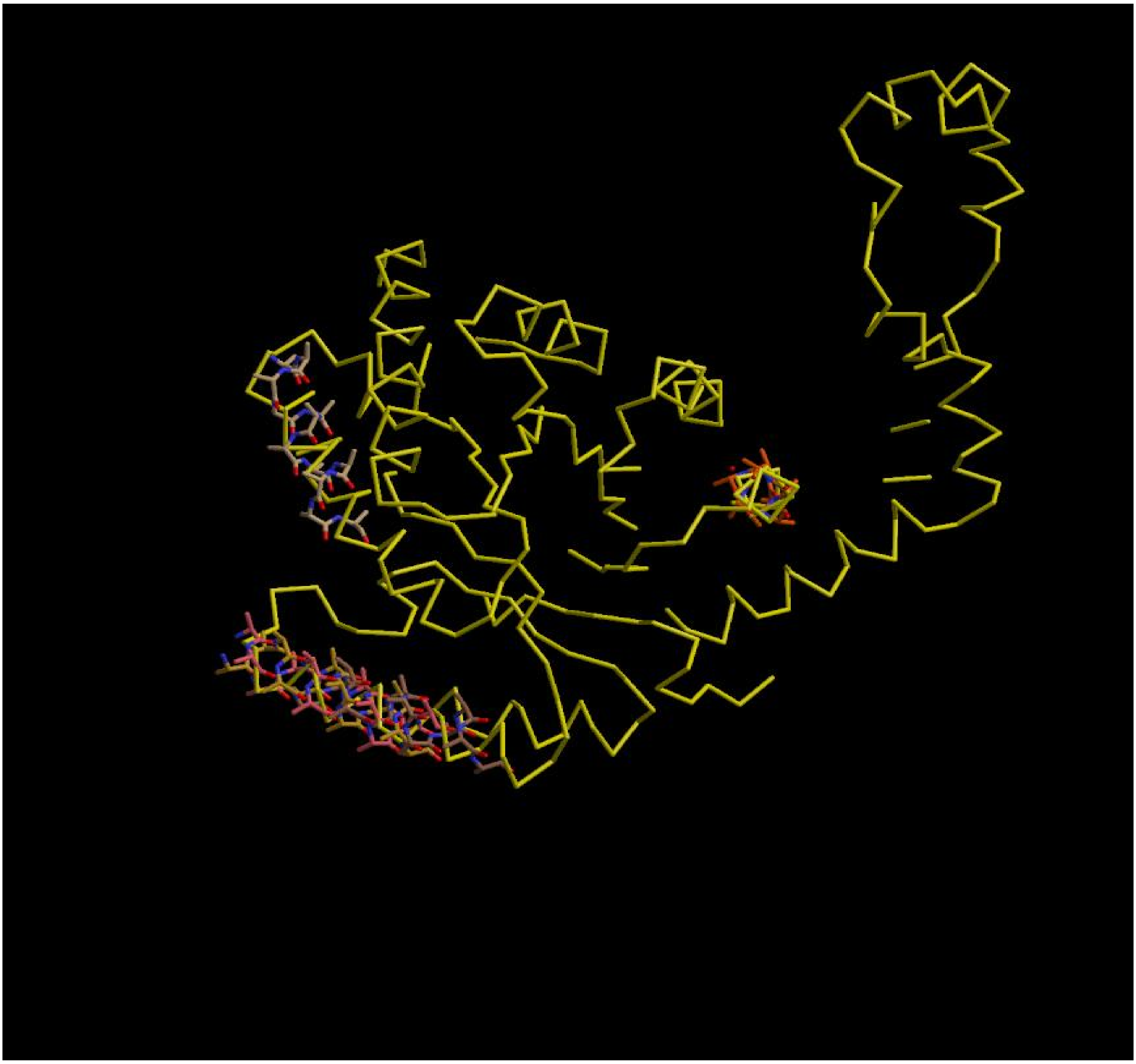
Translation search using the full resolution (1.54 Å) of the rotation solutions from data truncated to 2.1 Å leads to the correct solution, further improved by rigid body refinement. Nevertheless, this experimental substructure does not succeed in solving the structure with the equivalent SHELXE run rendering a solution with a MPE stuck at 71º whereas the incorporation of data extrapolation to 1 Å and iterative autotracing accomplishes final phases with an MPE of 19.1º. So, although it is clear that a very small fraction of a structure may suffice to phase it, it appears to be strongly dependent on how perfect this substructure is. In general, more than one fragment will need to be located to make up for the deviation from the real structure, which cannot be known *a priori*. In principle, it would be possible to do a brute-force generation of substructures[21] and test their expansion with SHELXE in all cases but the need to locate several fragments simultaneously strongly increases the number of parameters and thus the calculation time required in such an approach.

In the case of GI, the search for ideal model helices represents a more realistic scenario. Looking for $(Ala)_{12}$ helices, the first rotation search yields a high number of orientations, but all the highest solutions cluster around the same three helices in the model (see Supplementary Figure 1). Selecting the helix cluster that corresponds to the 18 aminoacids long helix comprising residues 64-81 in the model, rescoring the LLG of the rotation solutions (LLG=2.40 for the best solution) on longer helices of 16 aminoacids brings the maximum value up to LLG=3.29. For $(Ala)_{18}$ LLG further increases to 3.66 whereas for $(Ala)_{20}$ it drops to 3.44. This apparent sensitivity of the scoring function to the correctness of the fragment is worth exploiting as the length of straight helix one is looking for is not known until the structure is solved. No satisfactory translation solutions could be found for the $(Ala)_{18}$ helix, which further indicates the necessity of using the rotation function to improve the search models, making them closer to the real structure by introducing curvature.

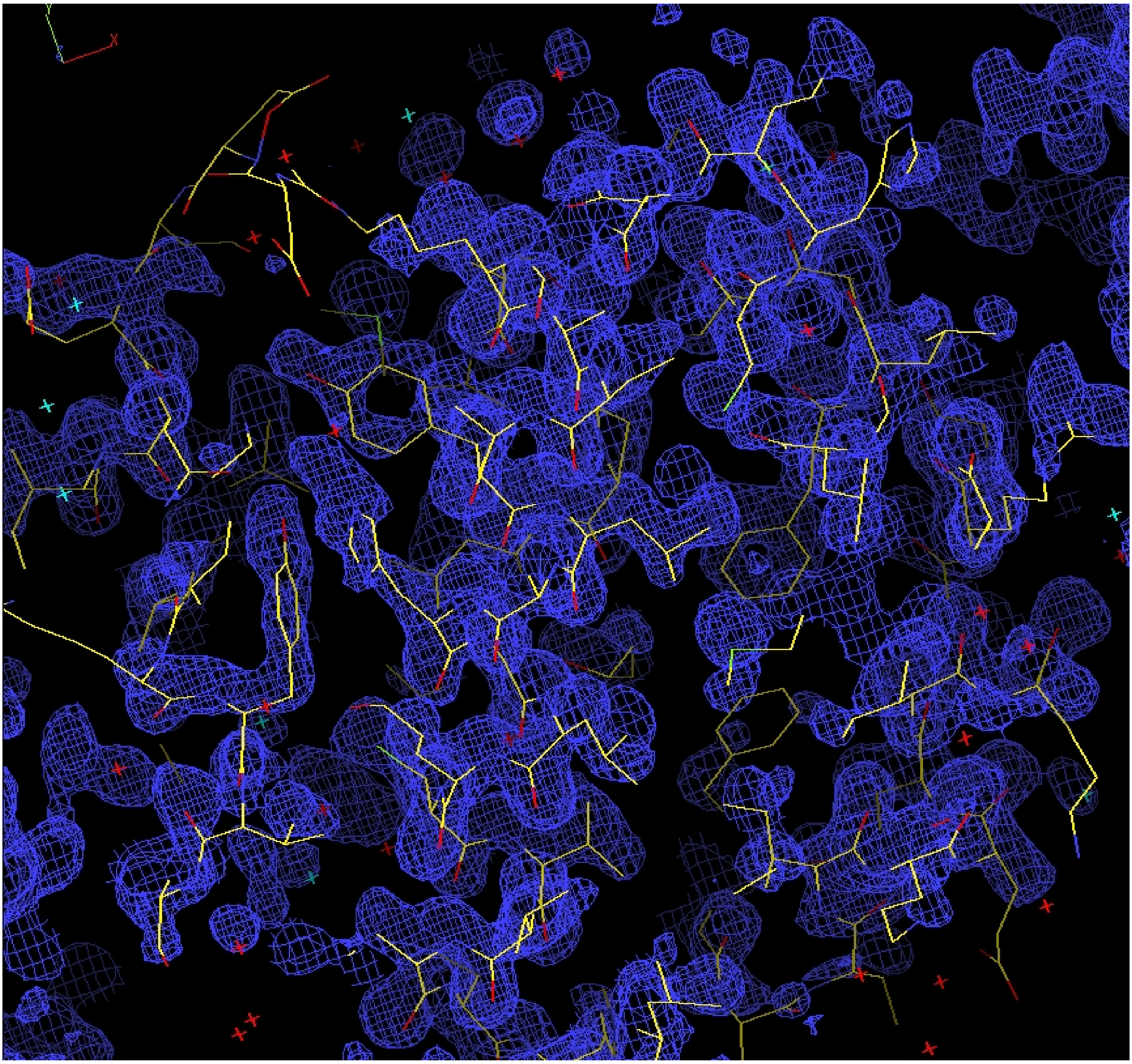
1.4. EIF5, solution of a mainly helical, 179 amino acids protein at 1.7 Å

The structure of EIF5 was solved using synchrotron data collected to a resolution of 1.69 Å, by the automated procedure setup with the strategy and parameters derived from our previous exhaustive tests. As the structure was already determined, we knew it to contain 11 alpha-helices of various lengths ranging from 7 to 21 amino acids and different degrees of deviation from the ideal straight helix. So the fragment search was set up to locate 7 helices of 12 alanines, restricting the resolution for the rotation search to 2.1 Å and using the full 1.7 Å resolution for the translation search and rigid body refinement with PHASER. The rotation search was carried out in 3º steps and translation in 0.7 Å steps. Combining every rotation peak with every translation peak would soon lead to an unmanageable number of solutions. In order to control the flow of the search, the following limits were set: For every rotation or translation search, peaks under 75% of top were rejected, as is the default in PHASER. Furthermore, from each translation run after the first fragment, no more than 70 solutions were further pursued. After the packing check, surviving substructures were divided in groups of 800 solutions for rigid body refinement and pruning of duplicates. Only the 100 top solutions were kept. Expansion to the full structure with SHELXE was attempted with substructures made up of 3, 4 and 5 helices. The structure was solved starting from 4 helices (240 atoms), yielding a MPE of 31º against the deposited model. Supplementary Figure 2 displays the resulting electron density map and the coordinates of the deposited model. In every SHELXE attempt, starting from phases derived from the partial structure calculated to a resolution of 1 Å, 4 runs of density modification made up of 20 cycles each were interspersed with autotracing. The density sharpening parameter (v) was set to 0 and reflections were extrapolated to a resolution of 1 Å. Substructures of 5 fragments also lead to an equivalent solution whereas no solution was achieved starting from 3 helices (180 atoms) even though the number of SHELXE iterations was doubled. The ARCIMBOLDO procedure was stopped beyond the 5th fragment.

Supplementary Figure 1 Location of $(\text{Ala})_{12}$ helices for the structure of glucose isomerase (yellow).



Supplementary Figure 2 Ab Initio Fo.fom electron density map for EIF5 derived from 4 ideal polyalanine helices of 12 residues after density modification. The deposited model is shown as well.



Nature Methods: doi:10.1038/nmeth.1365

Supplementary literature

19 Alexopoulos, E., Küsel, A., Sheldrick, G.M., Diederichsen, U. & Usón, I. (2004), *Acta Crystallogr*. **D60**, 1971-1980.

20 Morris, R.J. & Bricogne, G. (2003), Acta Crystallogr. **D59**, 615-617.

21 Glykos, N.M. & Kokkinidis, M. (2003), *Acta Crystallogr*. **D59**, 709-718.