Crystallographic *ab initio* protein structure solution below atomic resolution

Dayté D Rodríguez[1,5], Christian Grosse[2,5], Sebastian Himmel[3], César González[1], Iñaki M de Ilarduya[1], Stefan Becker[3], George M Sheldrick[2] & Isabel Usón[1,4]

***Ab initio* macromolecular phasing has been so far limited to small proteins diffracting at atomic resolution (beyond 1.2 Å) unless heavy atoms are present. We describe a general *ab initio* phasing method for 2 Å data, based on combination of localizing model fragments such as small α-helices with Phaser and density modification with SHELXE. We implemented this approach in the program Arcimboldo to solve a 222-amino-acid structure at 1.95 Å.**

Crystallography provides a view into the three-dimensional structure of biological macromolecules that is unsurpassed in detail and precision by any other structural technique. Nevertheless, the structural model product of the crystallographic analysis cannot be directly calculated from the experimental data and may rely to a large extent on interpretation owing to what is known as the phase problem: only the diffracted intensities and not the phases are determined from the X-ray diffraction experiment, whereas the phases are key to structure determination. In practice, initial phases can be derived either by molecular replacement with a related structure if available at all, with the drawback of introducing model bias, or from measurement of derivatives, which may result in an increase in experimental effort and time scale of the crystallographic study, as many derivatives are unsuccessful. Direct *ab initio* phasing of macromolecules, using only a dataset of native amplitudes without previous detailed structural knowledge or measurement of heavy atom or anomalous scatterer derivatives, was impossible until the advent of the dual-space algorithm[1]. This method, heavily relying on atomicity constraints, by exploiting mathematical conditions derived of the electron density being concentrated at randomly distributed, resolved, equal atom positions, is limited to those rare cases in which the protein crystal diffracts to around 1.0 Å resolution.

For small molecules, direct methods are almost invariably effective in solving crystal structures[2] but for macromolecules, both the

Table 1 | Summary of data for test proteins

Structure	PDB entry	Space group	Resolution for rotation (for translation)[a] (Å)	Number of residues (atoms)	Number of atoms (percentage of structure used for solution)
CopG	2CPG	$C222_1$	1.2	129 (1,015)	
Fragment Ala_{10} model helix			1.2 (1.2) 2.5 (2.5) 2.1 (1.5) 2.1 (1.8) 2.1 (2.1) 1.8 (2.1)		50 (5%) No solution 100 (10%) No solution No solution 100 (10%)
Oxidized bacteriophage T4 glutaredoxin (thioredoxin)	1ABA	$P2_12_12_1$	1.45	87 (728)	
Fragment Ala_{10} model helix			2.1 (1.45)		50 (7%)
Glucose isomerase	1MNZ	I222	1.54 in-house	385 (3,433)	
Perfect fragment[b] (amino acids 150–172)			2.1 (1.54) 1.54 (–)[c] 2.6 (–)		186 (5%) Rotation failed Rotation failed
Fragment Ala_{12}–Ala_{18} model helix			2.1		Rotation worked
Eukaryotic translation initiation factor 5 C-terminal domain (EIF5)	2IU1	$P2_12_12_1$	1.7	179 (1,473)	
Fragment Ala_{12} model helix			2.1 (1.7)		240 (16%)

[a] SHELXE expansion always with full resolution. [b] Perfect fragment with side chains cut out from final deposited model. [c] If rotation failed, translation cannot be attempted. After successful rotation if translation was unsuccessful in all attempts, no resolution is given.

[1] Instituto de Biología Molecular de Barcelona, Barcelona Science Park, Barcelona, Spain. [2] Lehrstuhl für Strukturchemie, Universität Göttingen, Göttingen, Germany. [3] Max Planck Institute for Biophysical Chemistry, Department of NMR-based Structural Biology, Göttingen, Germany. [4] Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [5] These authors contributed equally to this work. Correspondence should be addressed to I.U. (uson@ibmb.csic.es).

RECEIVED 3 APRIL; ACCEPTED 1 JULY; PUBLISHED ONLINE 16 AUGUST 2009; DOI:10.1038/NMETH.1365

larger number of atoms and the lack of atomic resolution hinder structure solution. *Ab initio* methods have been restricted to a few favorable cases, such as proteins of ~1,000 atoms diffracting to atomic resolution[3], or somewhat larger proteins diffracting to slightly more modest resolution if the structure contains heavy atoms (1.92 Å for 1,283 atoms including holmium and 1.65 Å for 7,890 atoms including 8 gold atoms)[4]. Of these two barriers preventing *ab initio* solution of macromolecular structures, resolution has proven to be the more difficult to overcome. Even extrapolation of nonmeasured data to fit partial phasing information has been shown to be more effective than leaving out missing data in experimental phasing[5]. Indeed, atomicity is a very powerful constraint, both in real and in reciprocal space. At lower resolutions, this constraint should be substituted by the knowledge that macromolecular structures are composed of smaller fragments of known geometry (that is, α-helices, β-strands and base pairs) and such fragments make up a first good approximation. Exploiting this fact to aid the phasing procedure, once a preliminary experimental map is available, is highly effective and has been implemented in the autotracing algorithms of programs such as RESOLVE[6], ARP/wARP[7] or SHELXE[8]. Tests have shown that correctly placed perfect fragments representing 13% of the structure can be enough for successful phasing through density modification with ACORN[9].

Here we present a generally applicable method to phase macromolecular structures from diffraction data with resolutions up to 2 Å. Our approach works in a multisolution frame by combining the location of small model fragments with density modification and autotracing of the resulting maps. *B*-factor refinement of the best resulting traces improves the interpretability of the map. After developing it on a number of previously determined test structures containing no atom heavier than sulfur, with resolutions of 1.2–2 Å (**Table 1** and **Supplementary Results**), we applied it to solve the previously unknown structure of the phosphotransferase system regulation domain II (PRD-II) from the transcriptional antiterminator protein GlcT of *Bacillus subtilis*[10]. This protein produced twinned crystals, constituting an additional hindrance. The crystals, containing 40% solvent and 222 amino acids in the asymmetric unit, diffracted to 1.95 Å.

We took great care to obtain the best possible experimental native data from a nonmerohedrally twinned crystal of PRD-II (twin ratio, 0.7:0.3) (**Supplementary Table 1**). The anomalous signal derived from the three sulfur atoms contained in the sequence was very weak and could not be exploited; the structure could not be solved by single-wavelength anomalous diffraction. The protein is a five-helix bundle, composed of 111 amino acids. It forms a dimer that deviates noticeably from twofold symmetry. Thus, it was not possible to determine beforehand whether the asymmetric unit contained 1 or 2 molecules so as whether to exploit noncrystallographic symmetry.

We searched for ideal α-helical polyalanine fragments of 14 residues with the program Phaser[11], truncating the resolution to 2.5 Å. As the fragments used in our method are very small (10–14 amino acids), they are accurate but represent a very low fraction of the total scattering mass. They can also fit the structure at many different nonequivalent positions. For instance, a 20-amino-acid helix might accommodate seven helices of 14 amino acids relatively displaced to each other by one amino acid. Therefore, Phaser returns many solutions (49 for the first fragment, 274 for the second, 1,473 for the third, 939 for the fourth, 3,167 for the fifth, 2,507 for the sixth, 7,920 for the seventh), with similar figures of merit, none above a conclusive threshold. Solutions close to their true location were present, but indistinguishable from the rest. There is no dependable criterion to identify unequivocally the best partial solutions. Therefore, we sent all partial solutions both to a new round of fragment searching with Phaser, to produce partial models with one more fragment and to density modification expansion with a beta-test version of the program SHELXE

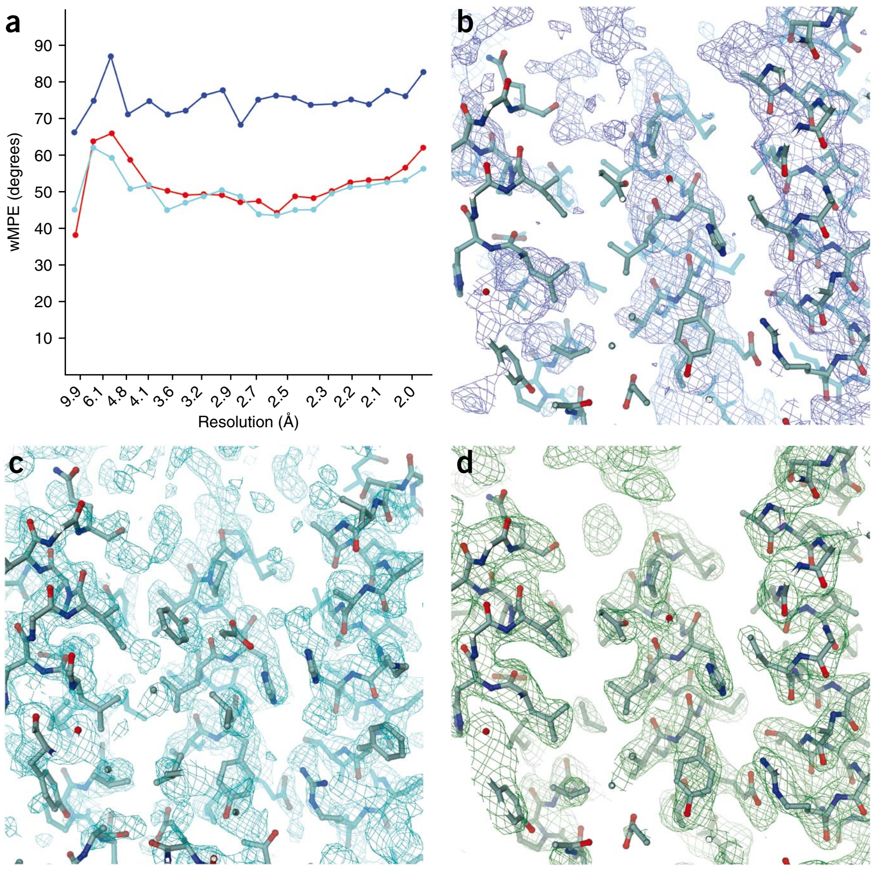

Figure 1 | Overall and detailed quality of the resulting phases. (**a**) Plot of mean phase error versus resolution for the initial phases derived from three helical Ala$_{14}$ fragments, after density modification and mainchain autotracing with SHELXE and after *B*-value refinement, taking the final refined model as reference. wMPE, *F*-weighted mean phase error. (**b**) Phaser F_o (observed structure factor) × figure of merit electron density maps derived from a set of 3 helices. (**c**) SigmaA weighted $2F_o - F_c$ (calculated structure factor) electron density map after *B*-value refinement of the mainchain atoms traced by SHELXE. (**d**) SigmaA weighted $2F_o - F_c$ electron density map from final Refmac5 refinement. The final model is displayed. Crystallographic object-oriented toolkit (COOT)[15] was used for real-space refinement and manual building. **Figure 1b–d** was prepared using DINO (http://www.dino3d.org/).

incorporating main-chain autotracing (http://shelx.uni-ac.gwdg.de/SHELX/). The program Arcimboldo (http://chango.ibmb.csic.es/ARCIMBOLDO/) controls this procedure and incorporates solution assembly and selection. Despite this method being computationally expensive, it can be easily parallelized and run on a grid or a multiprocessor cluster. In our case, we resorted to a local grid of Linux computers running Condor[12] and although the search for 8 fragments was still running, results for the sets of three fragments yielding the first solution were available in less than one day. After running SHELXE, true solutions can be distinguished by two clear-cut figures of merit: the number of residues the program has been able to trace and the correlation coefficient of the partial structure against the experimental data. Solutions can be then improved by *B*-value refinement of the main-chain atoms traced by SHELXE, with the program Refmac5[13], yielding even more easily interpretable electron density, although the global mean phase error (MPE) barely improved. This can be appreciated by plotting the weighted MPE versus resolution for the different phasing stages undertaken to solve the structure of PRD-II (**Fig. 1a**) and examining maps of the region around the C-terminal α-helix (chain A residues 96–111) calculated from the three helices located (main chain of 42 residues) (**Fig. 1b**), after iterative density modification and autotracing with SHELXE and after *B*-value refinement of the traced atoms with the program Refmac5 (**Fig. 1c**) and for the final model (**Fig. 1d**).

We obtained the structure solution for PRD-II after localizing the third fragment in three out of the 1,473 trials, and phasing became increasingly effective upon localizing additional correct fragments. This implies that phasing a 1,700-atom structure at 1.95 Å starting from 210 atoms (barely 12% of the structure) is possible. The MPE derived from the fragments leading to the best solution is 75.6°. The following density modification procedure with SHELXE was effective in lowering the MPE to 52.3°. First, we alternated 20 cycles of density modification three times with main chain autotracing initiated by a search for α-helical heptapeptides and common tripeptides in the resulting electron density maps. Finally, we performed a fourth 20-cycle density modification, incorporating data extrapolation beyond the actual experimental resolution limit of 1.95 Å up to 1.7 Å. We obtained extrapolated structure factors and phases by Fourier transformation of the density after each density modification cycle, followed by scaling the structure factors to fit an extrapolated Wilson plot[14]. *B*-value refinement of the main-chain fragments traced did not provide a global phase improvement but enhanced the interpretability of the electron density corresponding to the side chains, thus enabling the sequence to be assigned. Despite the apparently high MPE of the best solution, it could be pushed into a full solution with some manual intervention, with the possibility for automation through side-chain autotracing. If allowed to run, the Phaser procedure was also effective in locating the 8 longer helices in the structure, but the success with three helices shows that the method does not require the majority of the structure to be helical, and it is thus generally applicable at lower resolution (2 Å or better).

Feeding all 8 Phaser-determined helices into other conventional iterative model building, density modification and refinement programs was not effective in solving the structure of PRD-II. Molecular replacement with similar structures failed as well. From the available diffraction data, the structure of PRD-II could not be solved by any conventional method. Therefore, our general *ab initio* approach, consisting of multisolution location of short model main chain fragments coupled to density modification, autotracing and *B*-value refinement of the traced atoms, allowed us to address the phase problem for proteins diffracting to 2 Å.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Protein Data Bank (PDB): 3GWH.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by the European Union Integrated Project Biocrystallography (X) on a highly integrated technology platform for European Structural Genomics, the Spanish Ministry of Science and Innovation (grant BIO2006-06653), Fonds der Chemischen Industrie, Deutsche Forschungsgemeinschaft (DFG-IRTG1422) and the Max Planck Society. D.D.R. thanks Consejo Superior de Investigaciones Científicas for grant I3PMas_08_00057 for the Master in Crystallography and Crystallization, Universidad Internacional Menéndez Pelayo. We are also grateful to Institut d'Estadística de Catalunya for its contribution to grid computing.

AUTHOR CONTRIBUTIONS

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

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Miller, R. *et al. Science* **259**, 1430–1433 (1993).
2. Karle, J. & Hauptman, H. *Acta Crystallogr.* **9**, 635–651 (1956).
3. Sheldrick, G.M., Hauptman, H.A., Weeks, C.M., Miller, R. & Usón, I. *International Tables for Macromolecular Crystallography* vol. F, (eds., M.G. Rossmann and E. Arnold) 333–345 (Boston, 2001).
4. Caliandro, R. *et al. J. Appl. Crystallogr.* **41**, 548–553 (2008).
5. Caliandro, R. *et al. Acta Crystallogr.* **D61**, 556–565 (2005).
6. Terwilliger, T.C. *Acta Crystallogr.* **D59**, 38–44 (2003).
7. Ng, E.S., David, R.P., Azzola, L., Stanley, E.G. & Elefanty, A.G. *Nat. Struct. Biol.* **6**, 458–463 (1999).
8. Sheldrick, G.M. *Z. Kristallogr.* **217**, 644–650 (2002).
9. Jia-xing, Y., Woolfson, M.M., Wilson, K.S. & Dodson, E.J. *Acta Crystallogr.* **D61**, 1465–1475 (2005).
10. Schmalisch, M.H., Bachem, S. & Stülke, J. *J. Biol. Chem.* **278**, 51108–51115 (2003).
11. McCoy, A.J. *et al. J. Appl. Crystallogr.* **40**, 658–674 (2007).
12. Tannenbaum, T., Wright, D., Miller, K. & Livny, M. in *Beowulf Cluster Computing with Linux* (ed., T. Sterling) 307–350 (MIT Press, Cambridge, Massachusetts, USA, 2002).
13. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. *Acta Crystallogr.* **D53**, 240–255 (1997).
14. Usón, I., Stevenson, C.E.M., Lawson, D.M. & Sheldrick, G.M. *Acta Crystallogr.* **D63**, 1069–1074 (2007).
15. Emsley, P. & Cowtan, K. *Acta Crystallogr.* **D60**, 2126–2132 (2004).

ONLINE METHODS

Test cases of known structure. Four test cases were selected in different space groups at resolutions ranging from 1.2 to 2 Å and protein sizes from 87 to 368 amino acids in the asymmetric unit.

CopG (PDB code 2CPG): 1.2 Å, three molecules with 43 amino acids each, 129 amino acids, 1,015 atoms strand, helix, helix, 45% solvent, C222$_1$ (ref. 16).

Oxidized bacteriophage T4 glutaredoxin (thioredoxin) (1ABA): 1.45 Å, one molecule with 87 amino acids, 728 protein and 152 ligand plus solvent atoms, fold contains three helices and 4 strands, 45% solvent, P2$_1$2$_1$2$_1$ (ref. 17).

Glucose isomerase (1MNZ) 1.54 Å, in house data, 385 amino acids, TIM barrel, 3,433 atoms, 48% solvent, I222 (**Supplementary Fig. 1**).

EIF5 C-terminal domain (2IU1): 1.7 Å, one molecule with 179 amino acids, alpha-helical, 1,473 atoms, 45% solvent, P2$_1$2$_1$2$_1$ (ref. 18) (**Supplementary Fig. 2**).

Details on the data and tests are summarized in **Table 1** and in **Supplementary Results**.

Coordinates and observed structure factor amplitudes can be obtained on request for the structures where test data are not deposited (that is, in-house data from glucose isomerase).

Structure solution of PRD-II at 1.95 Å. The diffraction pattern consisted of two overlapping reciprocal lattices that could be cleanly indexed with two orientation matrices using the program CELL-NOW (Bruker AXS). Data were collected in four 180 degrees omega and one 360 degrees phi scans, taking advantage of the three-circle geometry to ensure good scaling. The data were integrated as a twin using the program SAINT (Bruker AXS) and scaled and reduced to a single unique dataset using TWINABS (Bruker AXS). 99,947 reflections were assigned to domain 1, 98,274 to domain 2 and 25843 to both domains. The twin volume ratio refined to 0.729:0.271 and the variation of the scale factors of the individual components with scan angle indicated that two individual crystals were present rather than an interpenetrant twin. The *R* factor for the agreement of the measured single and composite intensities with the values calculated from the unique reflection intensities and twin ratio was 0.091. The 12,803 unique data were 99.6% complete and had an effective redundancy of 15.5 (**Supplementary Table 1**).

Solution with Arcimboldo. The structure was expected to contain a four helix bundle, given its homology to the transcription antiterminator LICT (PDB code 1TLV). Molecular replacement with the standard programs using a search model derived from this set of coordinates was not successful. The length of the search helices was derived from this search model, containing 4 rather straight helices with more than 14 amino acids each. The unit cell dimensions were compatible with either one molecule and 60% solvent or two molecules and 30% solvent. Given the diffraction properties of the crystal, the lower solvent content appeared more probable, although the self-rotation function did not reveal a peak that would have settled the question. The fragment search was set up to locate 8 helices of 14 alanines, restricting the resolution for the rotation search, translation search and rigid body refinement with Phaser to 2.5 Å. This is Phaser's default and as the resolution question is not clearly settled and the optimum may vary from structure to structure we adopted it for a preliminary run as limiting the resolution results in shorter computation time. For each fragment, a rotation search followed by a translation search, a packing check and a rigid group refinement and clustering of solutions was carried out. The rotation search was carried out in 2° steps and translation in 0.7-Å steps. Solutions with clashes were discarded. 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 100 solutions were further pursued. After the packing check, surviving substructures were subject to rigid body refinement and pruning of duplicates. Expansion to the full structure with SHELXE was attempted with substructures made up of 2, 3, 4 and 5 helices. No solution was achieved starting from 2 helices (140 atoms). The structure was solved starting from 3 helices (210 atoms) in 3 of the 1,473 trial substructures (0.2% of the cases), starting from 4 helices (280 atoms) in 78 out of 939 cases and from the fifth fragment on, up to the eighth, the majority of the determined substructures lead to a solution. In every SHELXE attempt, starting from phases derived from the partial structure calculated to a resolution of 1.7 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 1.7 Å.

16. Gomis-Ruth, F.X. *et al. EMBO J.* **17**, 7405–7415 (1998).
17. Eklund, H. *et al. J. Mol. Biol.* **228**, 596–618 (1992).
18. Bieniossek, C. *et al. J. Mol. Biol.* **360**, 457–465 (2006).