Automated protein model building combined with iterative structure refinement

Anastassis Perrakis[1], Richard Morris[2] and Victor S. Lamzin[2]

In protein crystallography, much time and effort are often required to trace an initial model from an interpretable electron density map and to refine it until it best agrees with the crystallographic data. Here, we present a method to build and refine a protein model automatically and without user intervention, starting from diffraction data extending to resolution higher than 2.3 Å and reasonable estimates of crystallographic phases. The method is based on an iterative procedure that describes the electron density map as a set of unconnected atoms and then searches for protein-like patterns. Automatic pattern recognition (model building) combined with refinement, allows a structural model to be obtained reliably within a few CPU hours. We demonstrate the power of the method with examples of a few recently solved structures.

Since the beginning of protein crystallography, the need to improve and automate the structure solution steps has been a major focus of the field. Knowledge in most areas of modern molecular biology is accumulating at an accelerated rate. Scientists are seeking answers to a growing number of challenging biological questions and are thus determining the structures of large numbers of proteins and their complexes with substrates, inhibitors, other proteins and nucleic acids. The number of new structural entries deposited in the Protein Data Bank[1] has been increasing exponentially in the last few years and in 1997 alone reached ~ 1,400 structures. Thus the availability of fast, reliable, objective and easy-to-use procedures for building and refinement of structural models is becoming increasingly important. In view of the forthcoming era of structural genomics (proteomics), which is rapidly developing as a major and challenging area of structural biological research, the need for automated methods for solving macromolecular structures is now more pronounced than ever.

Assuming that the first bottleneck in structure determination, the expression and crystallization of the protein or macromolecular complex, is bypassed, the next most time-consuming step becomes the process of structure solution, model building and refinement. Recent instrumentation developments[2] permit an increasing fraction of the crystallographic community to access state-of-the-art synchrotron facilities in an almost routine manner, thus revolutionizing the field of diffraction data collection. Data to higher resolution and of better quality are more easily collected than ever, and experimental techniques for phase determination, such as multiple-wavelength anomalous dispersion (MAD[3]), are much more feasible. With the wealth of software available to optimize use of the diffraction data in obtaining phases, by heavy-atom methods (for example, Phases[4], MlPhare[5] and, more recently, SHARP[6], Solve[7] and CNS[8]), molecular replacement (for example, AMORE[9] and CNS[8]) or *ab initio* techniques[10–12], initial phases can be obtained more easily and more quickly than before. Subsequent use of phase improvement and extension techniques[13] can increase the quality of experimental maps.

However, a major time-consuming and critical step remains the construction of a molecular model to fit the electron density map. Especially in the case of maps of mediocre quality, this requires human intervention and laborious days at a graphics

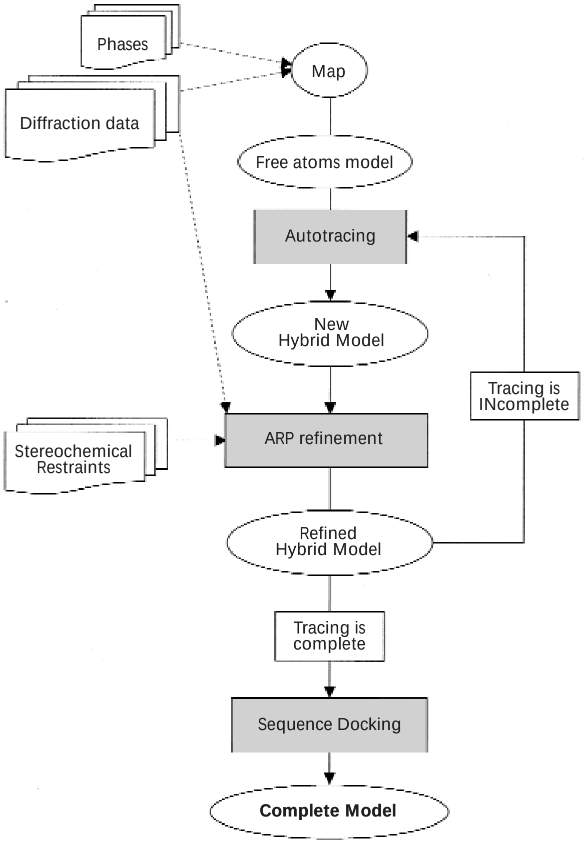

Fig. 1 A flowchart of the 'warpNtrace' procedure.

[1]European Molecular Biology Laboratory (EMBL), Grenoble Outstation, c/o ILL, BP 156, Av. des Martyrs, 38042 Grenoble, France. [2]EMBL, Hamburg Outstation, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany.

Correspondence should be addressed to A.P. *email: perrakis@embl-grenoble.fr*

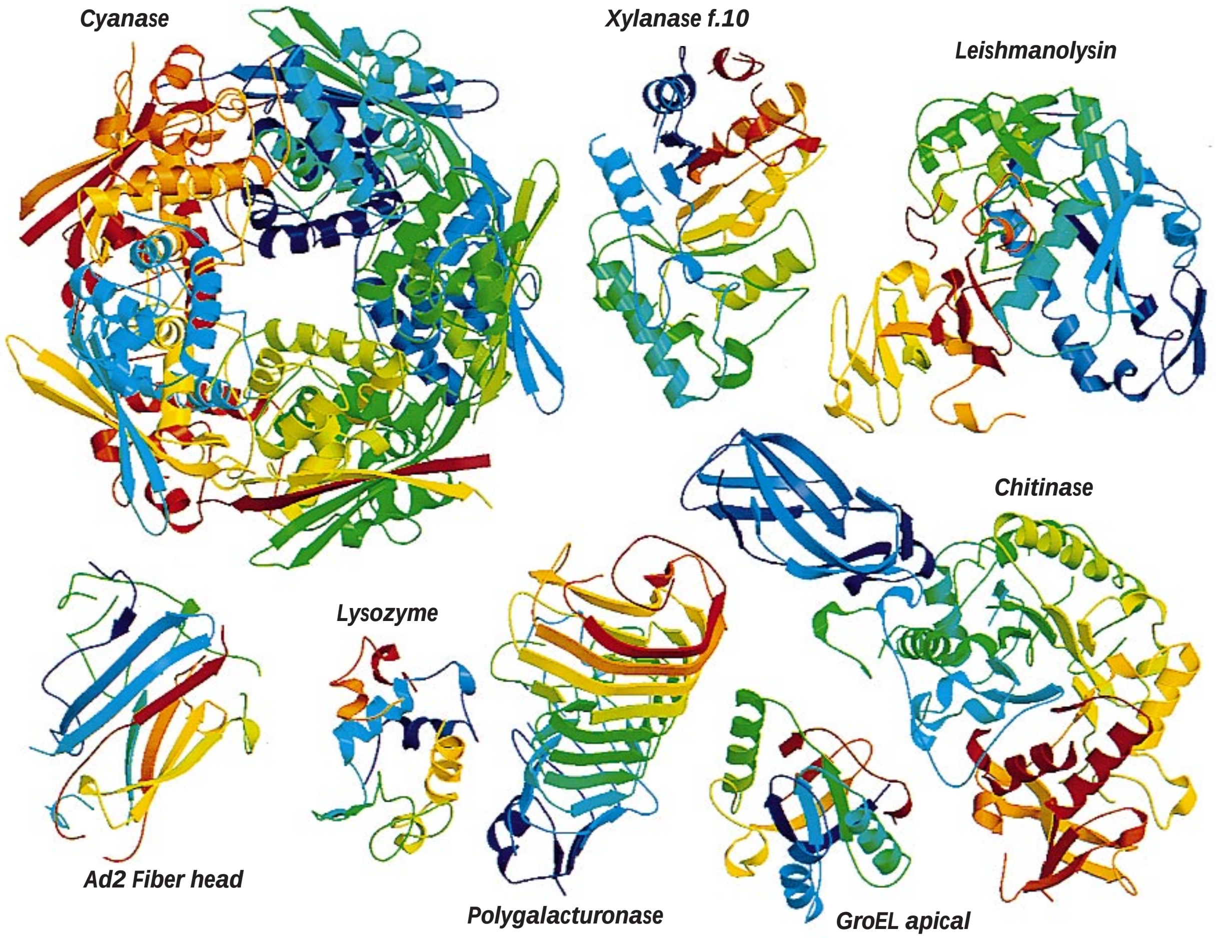

Fig. 2 A gallery of some of the structures automatically built and refined with warpNtrace. The method has no limitation on the size of the structure (compare cyanase with lysozyme), its fold (compare the chitinase $8\alpha/\beta$ barrel, the β-helix of polygalacturonase, the β-sheets in Ad2 fiber head and the $\alpha+\beta$ fold in others) or the number of domains (compare others with chitinase and leishmanolysin). The figures were drawn with Molscript[38], rendered by Raster3D[39], and compiled by The Gimp.

workstation. The initial model is often partial, and many cycles of refinement combined with further graphics sessions are required to approach a reliable model. The time required for this varies greatly depending on the experience of users and their familiarity with the software and the pitfalls of unrefined density models. An increasing number of new scientists, with a primary scientific focus on a biological problem, seek a quick and effortless structure solution. Consequently, the traditionally time-consuming (yet rewarding) procedure of unraveling the features of the electron density and expressing them as a stereochemical model of the macromolecule, calls for effectiveness, reliability and automation. Despite appreciable attempts to automate model building[14–16] as far as possible, procedures available to date rely heavily on the quality of the initial density map, and all require interactive decision making by the user.

The 'warpNtrace' procedure, which we describe here, is the first that automatically builds a protein model starting from electron density maps without user intervention. In addition, since by its nature it is coupled with refinement of the model and phases, as it proceeds it improves the quality of the map and also refines the automatically built model in both real and reciprocal space. The starting point can be any source of phase information: experimental, computational or a combination of these. An important requirement is that diffraction data extending to sufficient resolution (higher than 2.3 Å) are available. The time required for building a protein structure can then be shortened from several days or even months to a few CPU hours on inexpensive workstations.

Results

The 'warpNtrace' concept and related algorithms are described in detail in the Methods, and a flowchart is presented (Fig. 1). A few examples of applications are listed in Table 1, and a gallery of some of the structures solved is presented in Fig. 2. One example is extensively discussed, and six others are briefly considered.

Leishmanolysin. The structure of the *Leishmania* surface protein (Leishmanolysin, PSP) was originally solved with the use of SIRAS phases for two different crystal forms, multicrystal averaging, solvent flattening and density skeletonization[17]. In the present example we used the wARP phases derived as explained in ref. 18. In brief, a set of SIRAS phases was used extending to a resolution of 3.0 Å, which was extended to 2.5 Å by the CCP4[19] version of DM[20]. These phases were further extended by the wARP procedure to 2.0 Å resolution. The wARP model with the lowest R-factor was used to initiate model building. In the initial tracing, 252 residues were identified, belonging to 20 different main-chain fragments.

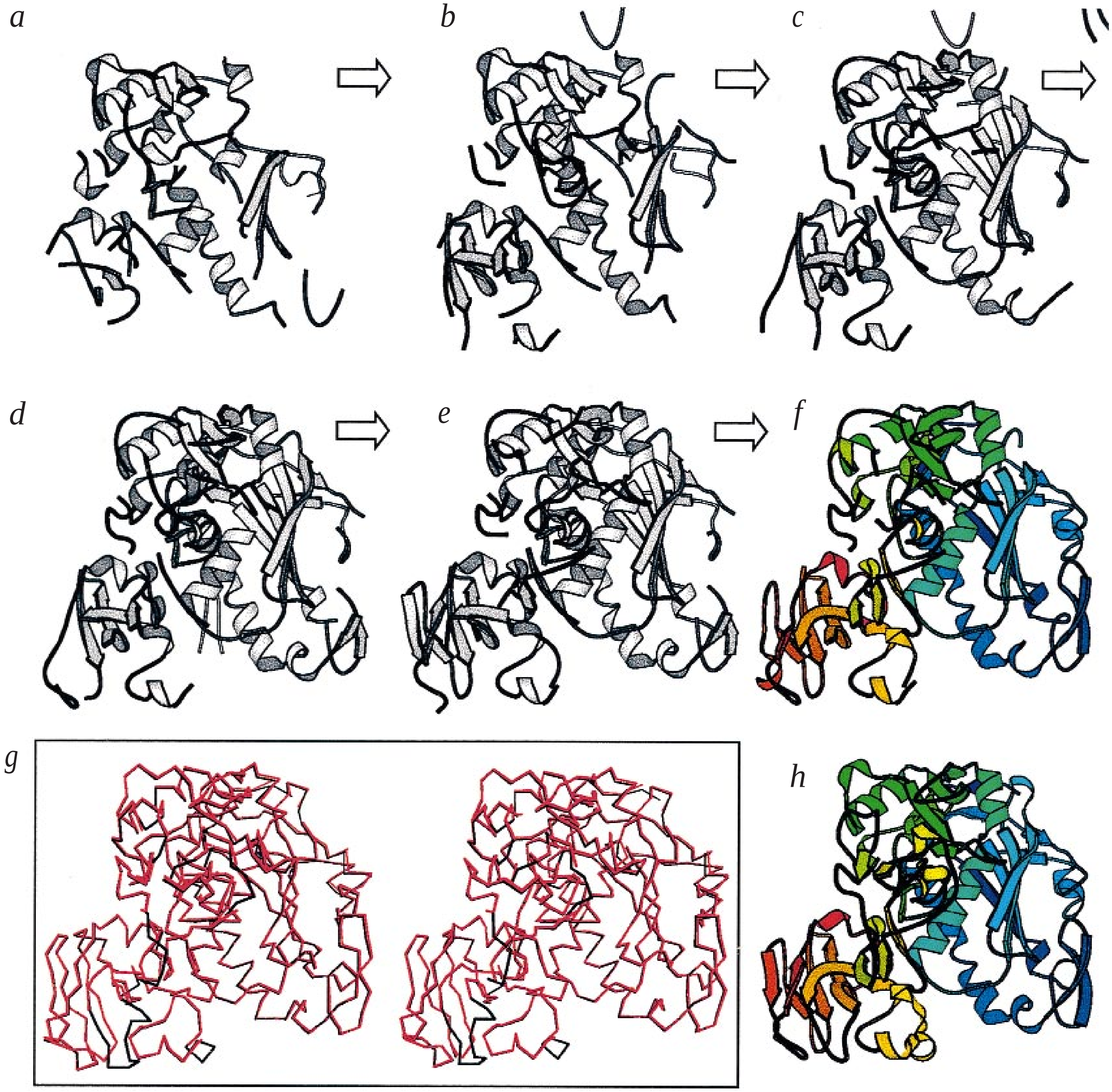

Fig. 3 Automatic building of Leishmanolysin. ***a,*** The model after the first autobuilding, and after ***b,*** 3, ***c,*** 6, ***d,*** 9, ***e,***12 and ***f,*** 15 cycles of autobuilding and refinement. Essentially the whole missing domain (bottom left of each panel) was completely recovered. ***g***, A stereo superposition of the autotraced (red) and the final (black) Cα trace. ***h***, a cartoon of the final model. Rainbow colors in (*f, h*) vary from the N- to the C-terminus. The figures were drawn with Molscript[38].

After autobuilding, 10 cycles of restrained ARP[21] were run, according to the standard protocol. One REFMAC[22] cycle of conjugate-gradient minimization to optimize a maximum-likelihood[23] residual was executed, applying the full calculated shifts and bulk solvent scaling. σ_A-weighted maps[24] were calculated and ARP was used to update the model. All atoms (main-chain, side-chain and free atoms) were allowed to be removed and new atoms were added where appropriate, as judged by ARP. After 10 iterations, a new building cycle was invoked. After every 'big' cycle a more complete model was available (Figs 3*a*-*f*, 4). This 'big' cycle was iterated 15 times. At the end, 450 residues were traced in seven chains. The longest chain, close to the N-terminus, contained 293 residues. All chains were docked unambiguously into the sequence, and 247 side chains were completely built. The autobuilt model has no differences from the final model in the traced part and virtually no errors; the r.m.s. displacement from all atoms of the final structure (Fig. 3*g*) is 0.28 Å, quite close to the expected coordinate error. The whole procedure took ~ 6 h, on either a Silicon Graphic R-10,000 or a 350 MHz 586 PC under Linux.

Cephalosporin synthase. This structure was solved from merohedrally twinned crystals[25]. The native data set used for running 'warpNtrace' was only approximately de-twinned, as there was no way to refine the twinning fraction. Nevertheless, most of the structure could be automatically built. The relatively low quality of the building (considering resolution) can be attributed to the presence of the crystal twinning.

β-Mannanase. Because the sequence of the protein was not known when the first maps were available, the autotraced fragments with the highest score for the sequence assignments were used to design oligonucleotide primers to clone the gene as described[26], exemplifying the combination of structural and genetic approaches in modern molecular biology.

Cyanase. Potential difficulties with this, the largest structure built so far with ARP/wARP 'warpNtrace', include the presence of 10 monomers in the crystallographic asymmetric unit, the formation of the decamer by a pentamer of dimers, and the unusual way that dimers themselves are formed (M.A. Walsh and co-workers, manuscript in preparation). Building 1,560 amino acids into the electron density would take some time, but 'warpNtrace' completed the task overnight.

α-Adaptin 'ear' domain. Considering the resolution (1.9 Å) and the excellent quality of the data and starting map (D.J. Owen, P.R. Evans and co-workers, manuscript in preparation), this is one of the most remarkable 'failures' of 'warpNtrace'. Although at this resolution an essentially complete trace was expected, only 196 out of 238 residues were autotraced and 44 side chains were autobuilt. The breaks in the tracing correspond to flexible loops, which could, however, be built manually. Investigating the reasons for the relatively poor performance of 'warpNtrace' in this case may lead to further improvements of the procedure.

GroEL apical domain. In one of the most exciting examples of a modern crystal structure determination, the MAD data for phasing this structure were collected within 20 min, as described by M.A. Walsh and co-workers (manuscript in preparation). Since data processing, location of the heavy-atom sites and phasing were straightforward, subsequent phase extension and autobuilding using 'warpNtrace' delivered a complete (and almost fully refined) structure within a few hours. Whereas three days elapsed from the start of diffraction data collection to the protein structure, in retrospect the entire procedure could have been easily performed within 24 h.

Table 1 Summary of the structures autotraced with ARP/wARP to date[1]

Structures[1–3]	Resolution (Å) (data, phases)[4]	Initial phases (method, software)	Residues/ molecules	Number chains traced (a.u.)	Number residues traced (% of all)	Number of side chains (% of traced)[4]
Lysozyme (1)	0.9, N/A	Direct methods, shake'n'bake	128/1	1	126 (98%)	N/A
Endoglucanase CelA (2)	0.9, 0.9	SAD, MlPhare	380/1	2	354 (93%)	284(81%)
Rubredoxin (3)	1.2, N/A	Native Patterson, Shelxs	52/1	1	50 (96%)	36 (72%)
****Cephalosporin synthase*** (4)	1.3, 2.7	MIRAS, SHARP/Solomon	311/1	7	262 (84%)	N/A
HisF (5)	1.4, 1.9	MAD, SHARP	253/1	3	235 (93%)	N/A
***β-Mannanase** (6)	1.5, 2.4	MIRAS, SHARP/Solomon	302/1	6	292 (97%)	N/A
Ad2 fiber head (7)	1.5, N/A	MR, 68% identity, AMORE	195/1	3	188 (96%)	130 (69%)
Auracyanin (8)	1.5, 2.4	MAD Cu, MlPhare	138/1	2	135 (98%)	57 (50%)
Glutamate mutase ***Clostridia cochlearium*** (9)	1.6, 1.8	MIRAS, MlPhare	1,240 / 2 + 2	7	1216 (98%)	N/A
Sol. Lytic transglycosylase 35 (10)	1.7, 2.7	MIRAS, Phases	301/1	5	279 (93 %)	N/A
****GroEL Apical domain*** (11)	1.7, 2.5	MAD Se, MlPhare/DM	145/1	2	138 (95%)	74 (54%)
****Cyanase*** (12)	1.8, 2.0	MAD Se, CNS/DM	1,580/10	17	1,495 (95 %)	1,095 (73%)
Polygalacturonase (13)	1.8, 2.7	MIRAS, Phases	670/2	8	645 (96 %)	498 (77%)
P13 Kinase **P85-α SH2 domain** (14)	1.8, N/A	MR, NMR model, AMORE, Xplor	107/1	3	97 (91%)	68 (72%)
I-PpoI endonuclease (15)	1.8, N/A	MR, apo- form, AMORE/CNS	326/2	9	294 (90%)	N/A
Xylanase family 10 (16)	1.8, N/A	MR, X-PLOR	302/1	3	297(98%)	175 (59%)
*α-Adaptin 'ear' domain (17)	1.9, 1.9	MIRAS, SHARP,/Solomon	238/1	9	196 (82%)	44 (22%)
Xylanase family 11 Hg (18)	2.0, 2.0	SAD, SHARP/Solomon	200/1	4	191 (96 %)	92 (48%)
**Leishmanolysin* (19)	2.0, 2.5	SIRAS, MlPhare/DM	475/1	8	456 (96 %)	247 (55%)
L-Aspartate oxidase (20)	2.1, 2.3	MIR, SHARP/Solomon	480/1	5	370 (77 %)	350 (95%)
Stat-3 (21)	2.2, 2.5	MIRAS, SHARP/Solomon	580/1	22	295 (51 %)	N/A
*Chitinase A *Serratia marsescens* (22)	2.3, 2.5	SIRAS, Phases	536/1	9	515 (96 %)	269 (52%)

[1] The survey was done in the arp-users email list, eight weeks after the release of the software. All computer jobs took 6–12 h at a variety of state-of-the-art workstations. Structures marked with * are discussed in more detail in text.

[2] Structures with the protein name in *italics* were used as examples, tests and benchmarks for program performance. Structures in normal text were autotraced with ARP/wARP after they had been solved and traced by the corresponding authors. Structures in **bold** were traced by ARP/wARP while the users were already performing model building and were actually used in parallel as a reference. The ones in ***bold italic*** were used as the only model during structure solution.

[3] References for the listed structures are given in parentheses following the structure name, as follows: (1) see ref. 33; (2) P. Alzari; (3) see ref. 34; (4) see ref. 25; (5) D.A. Lang, G. Obmolova, R. Thoma, R. Sterner & M. Wilmanns; (6) see ref. 26; (7) M. van Raaij & S. Cusack; (8) C.S. Bond & H.C. Freeman; (9) R. Reitzer; (10) see ref. 35; (11,12) M. Walsh & A. Joachimiac; (13) Y. van Santen & B. Dijkstra; (14) F.J. Hoedemaeker, G. Siegal, P.C. Driscoll & J.P.A. Abrahams; (15) M. Miller, B. De Latte & K. Krause; (16) see ref 36; (17) D.J. Owen & P.R. Evans; (18) Z. Dauter; (19) see ref. 17; (20) A. Mattevi; (21) see ref. 37; (22) see ref. 27

[4] N/A stands either for not applied or not applicable, depending on circumstance.

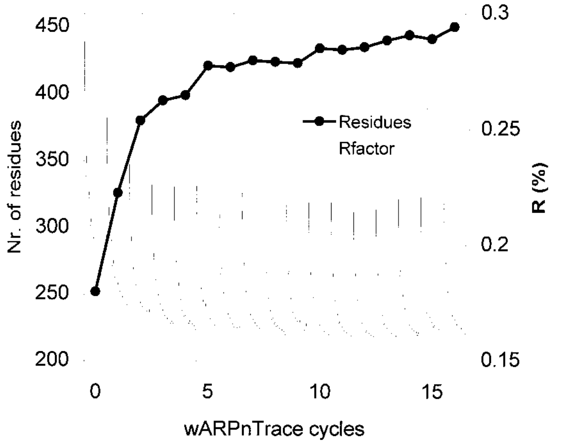

Fig. 4 Number of traced residues and the crystallographic R-factor as a function of warpNtrace cycles on the example of leishmanolysin. After each autobuilding, the initial R-factor is higher but quickly converges to a lower value. The model becomes gradually more complete.

Chitinase. This is the lowest resolution application to date. Surprisingly enough, the data are of mediocre quality[27] and 2.3 Å was the real resolution limit. However, the high solvent content (61%) enabled an average of seven observations per atom, easily allowing an almost complete trace.

Discussion

Resolution of the diffraction data. The native diffraction data should be of high resolution. In general, the number of X-ray reflections should be at least six to eight times higher than the number of atoms in the model. This roughly corresponds to a resolution of 2.3 Å for a crystal with 50% solvent content. However, the method can work with lower resolution or fail with a higher one, depending less on the quality of the initial phases and more on the internal quality of the data and on the inherent disorder of the molecule.

Quality of starting phases. Given that well-established algorithms can do initial phase extension to the resolution limit of the diffraction data, the quality of the starting phases is not a real limitation. Clearly, the higher the quality of starting phases, the less time is required for 'warpNtrace'. Our experience with the procedure has increased confidence to the point of being encouraged to try it on any map that appears to contain secondary-structure elements. A few minutes to launch the procedure and a few computer cycles will determine beyond doubt whether 'warpNtrace' can be successfully used to solve and refine a particular structure.

Quality of the diffraction data. The X-ray data should be complete, especially in the low-resolution range (5 Å and lower). During reciprocal space refinement a bulk solvent correction is essential. If the strong low-resolution data are systematically incomplete, then the density map, even in the case of a good model, is usually discontinuous; this can produce poor autotracing results.

Quality and validation of the model. Because of the strict criteria used for the automated model building, we have observed no errors in the main-chain tracing. However, we stress that the automatically built model should be inspected with care. Cross-validation by means of an R_{free}[28] is provided by this tool and can be routinely used throughout the procedure.

Applicability of the method. The current requirement of 'warpNtrace' that the diffraction data must extend to a resolution of at least 2.3 Å may seem a limitation. However, about two thirds of the crystal structures in the PDB (~5,000 entries) are determined at a resolution of 2.3 Å and higher, of which 3,000 were determined in the last three years alone. The number of potential applications of our procedure is directly linked to this tendency, and we expect the method to gain in applicability and importance.

Methods

Background. Our work on ARP (automated refinement procedure) for interpreting the experimental crystallographic density maps as sets of unconnected atoms (free-atom models), iterative updating of these sets and employment of averaging techniques to improve phase quality, has been described[18,21,29]. In brief, ARP is based on a cyclic procedure of fitting the calculated to the observed structure factor amplitudes in reciprocal space, followed by density-based addition and deletion of atoms in real space.

An overview of the method. The 'warpNtrace' procedure (Fig. 1) is based on the interpretation of the electron density map as a *hybrid* model consisting of a conventional protein model (chains of connected amino acids) and a set of free atoms (unconnected atoms of uniform atomic type), which are refined by ARP. Information from parts of the map and from the free-atom model can be automatically recognized to contain elements of protein structure, and at least a partial atomic protein model can be built. This model typically does not fully describe the electron density map, and hence a combination of this partial protein model with a set of free atoms (a *hybrid* model) allows a considerably better description of the current map. The protein model provides additional information (stereochemical restraints), while free atoms describe prominent features in the electron density, unaccounted for by the protein model. Given the extra information, in the traditional form of stereochemical restraints, which are essential for refinement to proceed[30,31], improved phases can be obtained and a better model can be constructed. These steps are iterated and the improved phases allow construction of even larger parts of the model, until an almost complete protein model is obtained in a fully automated manner (Fig. 3).

Main-chain autotracing. The main chain (backbone) of any protein consists of nonbranching, nonoverlapping chains of structurally identical Cα-C-O-N-Cα peptide units. To recognize such patterns in an electron density map, each atom of the *hybrid* or free-atom model is assigned a probability of being correct (a score), using the refined atomic displacement parameter and the height of the electron density at the atomic center. Pairs of atoms that have the highest score and are located within 3.8 ± 0.5 Å from each other are considered as possible successive Cα atoms. An ideal *trans* peptide is then fit between them in the two possible directions. The Cα-Cα pair is kept for further consideration only if the density for the carbonyl oxygen is above a certain threshold. All possible polypeptide chains that fit the expected conformations known from protein databases are constructed. This is done through a connectivity table of peptides that run in the same direction, share a common Cα atom and have a Cα-1 to Cα+1 distance within 4.6–7.8 Å . Branchpoints (that is, Cα atoms with more than two connections) are resolved by eliminating the peptide(s) in lower density. Subsequently the longest chain is accepted and chains spatially overlapping with it are eliminated. The next longest chain is then selected and the procedure iterated until no more chains longer than four peptides (five residues) remain. There are two main reasons that several backbone fragments rather than one are located: the probabilistic identification of peptide units and the naturally high conformational flexibility of connections between them. Combined with the often insufficient quality of X-ray data and/or phases, these introduce large enough errors to cause either density breaks (discontinuity) or density overlaps (branching). Only peptides traced without any ambiguity and chains with clearly resolved branching characteristics are considered. By using strict criteria for the autotracing, at the expense of introducing a few extra breaks and leaving a few ambiguous residues unassigned, we minimize possible bias.

Iteration. The *hybrid model* is subsequently refined in ARP cycles. Refinement using the new stereochemical data allows more accurate phases to be obtained, which give higher quality maps and allow more reliable and complete autotracing. During ARP refinement, atoms are removed whether they are free atoms or part of the protein model; new atoms are added where necessary. Moreover, on each cycle the previous model is discarded during autotracing, that is, all atoms are treated as free. Thus we minimize possible bias introduced by an incorrect interpretation of the density in early stages and imposed during refinement.

Side-chain identification and sequence docking. For any side chain, the connectivity between its atoms can be expressed as a vector, with the elements containing the number of connections; for example, alanine is '1', valine is '12', phenylalanine is '11221'. A similar vector is calculated for every autobuilt residue, showing the *observed* connectivity between its Cα and free atoms. The known protein sequence is expressed as a two-dimensional matrix built from the connectivity vectors for each amino acid. Each polypeptide fragment is given a similar 'observed' connectivity matrix. The 'observed' matrix is then slid along the sequence matrix and a score is calculated for each possible docking position, in a way similar to ref. 32. This then allows automated inspection of the side-chain densities, search for expected patterns, and building the most probable side-chain conformations.

Assembly of a globular molecule. The final step is the 'assembling' of the polypeptide fragments, together with the side chains built, into a globular molecule. More complex cases include resolving multiple copies of the molecule or distinct domains.

Acknowledgments

The authors are extremely grateful to K. Wilson and Z. Dauter for many useful comments and discussions. We also acknowledge the numerous test users for their efforts in using (and tuning) the ARP/wARP software and for making available brief results (presented in Table 1). Many thanks to our colleagues for critical reading of the manuscript. 'warpNtrace' is a part of the ARP/wARP suite, http://www.embl-hamburg.de/ARP (thanks to C. Blessing and T. Eriksson for the www setup and useful suggestions). The authors and the European Molecular Biology Laboratory copyright the software, which is available to the community.

Received 14 December, 1998; accepted 16 February, 1999.

1. Bernstein, F.C. *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542 (1977).
2. Helliwell, J.R. Synchrotron radiation facilities. *Nature Struct. Biol.* **5**, 614–617 (1998).
3. Ogata, C.M. MAD phasing grows up. *Nature Struct. Biol.* **5**, 638–640 (1998)
4. Furey, W. & Swaminathan, S. PHASES-95: a program package for the processing and analyzing diffraction data from macromolecules. *Methods Enzymol.* **277**, 590–620 (1997).
5. Otwinowski, Z, In *Isomorphous replacement and anomalous scattering. Proceedings of the CCP4 study weekend, January 25–26, 1991* (eds Wolf, W., Evans, P.R. & Leslie, A.G.W.) 80–85 (Daresbury Laboratory, Warrington, UK; 1991).
6. Fortelle de La, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 590–620 (1997).
7. Terwilliger, T.C. & Berendzen, J. Automated structure solution for MIR and MAD. *Acta Crystallogr. D* **55**, 849–861 (1999).
8. Brünger, A.T. *et al. Crystallography & NMR System*: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
9. Navaza, J. *AMoRe*: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
10. Miller, R., Gallo, S.M., DeTitta, G.T., Khalak, H.G. & Weeks, C.M. SnB, crystal structure determination via shake-and-bake. *J. Appl. Crystallogr.* **27**, 613–621 (1994).
11. Sheldrick, G.M. Patterson superposition and *ab initio* phasing. *Methods Enzymol.* **276**, 307–326 (1997).
12. Lunin, V.Y., Lunina, N.L., Petrova, T.E., Urzhumtsev, A.G. & Podjarny, A.D. On the *ab initio* solution of the phase problem for macromolecules at very low resolution. II. Generalized likelihood based approach to cluster discrimination. *Acta Crystallogr. D* **54**, 726–734 (1998).
13. Abrahams, J.P. & de Graaf, R.A.G. New developments in phase refinement *Curr. Opin. Struct. Biol.* **8**, 601–605 (1998).
14. Jones, T.A., Zou, J.-Y., Cowan, S.W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
15. Oldfield, T.J. A Semi-Automated map fitting procedure. In *Proceedings from the 1996 meeting of the International Union Of Crystallography Macromolecular Computing School* (1996).
16. Kleywegt, G.J. & Jones, T.A. Template convolution to enhance or detect structural features in macromolecular electron-density maps. *Acta Crystallogr. D* **53**, 179–185 (1997).
17. Schlagenhauf, E., Etges, R. & Metcalf, P. The crystal structure of *Leishmania major* surface proteinase leishmanolysin (gp63). *Structure* **6,** 1035–1046 (1998).
18. Perrakis, A., Sixma, T.K., Wilson, K.S. & Lamzin, V.S. wARP: improvement and extension of crystallographic phases by weighted averaging of multiple refined dummy atomic models. *Acta Crystallogr. D* **53**, 448–455 (1997).
19. Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
20. Cowtan, K. & Main, P. Miscellaneous algorithms for density modification. *Acta Crystallogr. D* **54**, 487–493 (1998).
21. Lamzin, V.S. & Wilson, K.S. Automated refinement of protein models. *Acta Crystallogr. D* **49**, 129–147 (1993).
22. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method**.** *Acta Crystallogr. D* **53**, 240–255 (1997).
23. Bricogne, G. Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. *Acta Crystallogr. D* **49**, 37–60 (1993).
24. Read, R.J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42**, 140–149 (1986).
25. Valegard, K. *et al.* Structure of a cephalosporin synthase. *Nature* **394**, 805–809 (1998).
26. Hilge, M. *et al.* High resolution native and complex structures of thermostable β-mannanase from *Thermomonospora fusca*—substrate specificity in glycosyl hydrolase family 5. *Structure*, **6**, 1433–1444 (1998).
27. Perrakis, A. *et al.* Structure of a bacterial chitinase at 2.3 Å resolution. *Structure* **2**, 1169–1180 (1994).
28. Brünger, A.T. Assessment of phase accuracy by cross validation: the free R value. Methods and application. *Acta Crystallogr. D* **49**, 24–36 (1993).
29. Lamzin, V.S. & Wilson, K.S. Automated refinement for protein crystallography. *Methods Enzymol.* **277**, 269–305 (1997).
30. Sussman, J.L., Holbrook, S.R., Church, G.M. & Kim, S.-H. A structure factor least-squares refinement procedure for macromolecular structures using constrained and restrained parameters. *Acta Crystallogr. A* **33**, 800–804 (1977).
31. Hendricksoon, W.A. & Konnert, J.H. In *Computing in crystallography* (eds Diamond, R., Ramasechan, S. & Venkatesan, K.), 13.01–13.25 (Indian Institute of Science, Bangalore, India; 1980).
32. Zou, J.-Y. & Jones, A. Towards the automatic interpretation of macromolecular electron density maps: qualitative and quantitative matching of protein sequence to map. *Acta Crystallogr. D* **52**, 833–841 (1996).
33. Deacon, A.M., Weeks, C.M., Miller, R. & Ealick, S.E. The shake-and bake structure determination of troclinic lysozyme. *Proc. Natl. Acad. Sci. USA* **16**, 9284–9289 (1998).
34. Dauter, Z., Sieker, L.C. & Wilson, K.S. Refinement of rubredoxin from *Desulfovibrio vulgaris* at 1.0 Å with and without restraints. *Acta Crystallogr. B* **48**, 42–59 (1992).
35. van Asselt, E.J., Perrakis, A., Kalk, K.H., Lamzin, V.S. & Dijkstra, B.W. Accelerated X-ray structure elucidation of a 36 kDa muramidase/transglycosylase using wARP. *Acta Crystallogr. D* **54**, 58–73 (1998).
36. Schmidt, A., Schlacher, A., Steiner, W., Schwab, H. & Kratky, K. Structure of the xylanase from *Penicillium simplicissimum. Protein Sci.* **7**, 2081–2088 (1998).
37. Becker, S., Groner, B. & Müller, C.W. Three dimensional structure of the Stat3β homodimer bound to DNA *Nature* **394**, 145–151 (1998).
38. Kraulis, P.J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
39. Merrit, E.A. & Bacon, D.J. Raster3D photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).