Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure elucidation of β-mannanase: from the electron-density map to the DNA sequence

Mark Hilge,[a]*† Anastassis Perrakis,[b] Jan Pieter Abrahams,[c] Kaspar Winterhalter,[a] Klaus Piontek[a] and Sergio M. Gloor[a]

[a]Institute of Biochemistry, Federal Institute of Technology, ETH Center, CH-8092 Zürich, Switzerland, [b]EMBL, c/o ILL, BP 156, F-38042 Grenoble CEDEX, France, and [c]Biophysical Structural Chemistry, Gorlaeus Laboratories, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands

† Present address: Biophysical Structural Chemistry, Gorlaeus Laboratories, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands.

Correspondence e-mail: hilge@chem.leidenuniv.nl

Received 31 July 2000
Accepted 27 October 2000

PDB Reference: β-mannanase, 1bqc.

The crystal structure of affinity-purified *Thermomonospora fusca* β-mannanase has been solved despite the lack of the major part of the amino-acid sequence. A high-quality electron-density map allowed the identification of a stretch of eight amino acids close to the C-terminus which was used to design a degenerate downstream PCR primer. Together with a specific primer previously derived from the N-terminus, 95.7% of the mannanase gene sequence was obtained from genomic *T. fusca* DNA by PCR. The structure-derived sequence was then compared with the DNA-derived sequence and corrected when necessary. Applying the presented protocol, there was no need to manually build a model at an early stage of structure determination, an erroneous and tedious process, especially in the absence of the amino-acid sequence. Using the DNA sequence information and the current version of *ARP/wARP*, 281 residues, or 93% of the polypeptide chain (including side chains), were built and refined to an *R* factor of 16.5% without any manual intervention.

1. Introduction

Despite the remarkable achievements in molecular biology, some proteins still behave very intractably towards cloning of their cognate DNA or cDNA. If they can be purified to homogeneity by biochemical methods, N-terminal and internal amino-acid sequence information can usually be obtained. However, this may not be sufficient to rapidly clone their gene by direct approaches comprising PCR and library screening. The reasons might be manifold: too short and ambiguous amino-acid stretches, high base degeneracy, minor impurities in the protein preparation leading to false sequence information derived from more than one protein, existence of highly related genes or very asymmetric base distribution of the cognate gene and possibly even the entire organism. Hence, other approaches are required to deduce the primary structure. Along with the clear focus on atomic structure analysis, protein crystallography can often provide the essential information to clone the target gene. Since the amino-acid sequence information derived from the electron-density map encompasses different regions of the same protein, it is possible to design appropriate DNA probes for various cloning strategies.

The β-mannanase from the thermophilic actinomycete *T. fusca* presents an example of such a protein. β-Mannanases hydrolyse the O-glycosidic bonds in mannan, a hemicellulose constituent of plants (McCleary, 1988). With a temperature optimum of 353 K and a broad pH tolerance, the *T. fusca* mannanase might be particularly useful in industrial processes performed under elevated temperature and alkaline pH

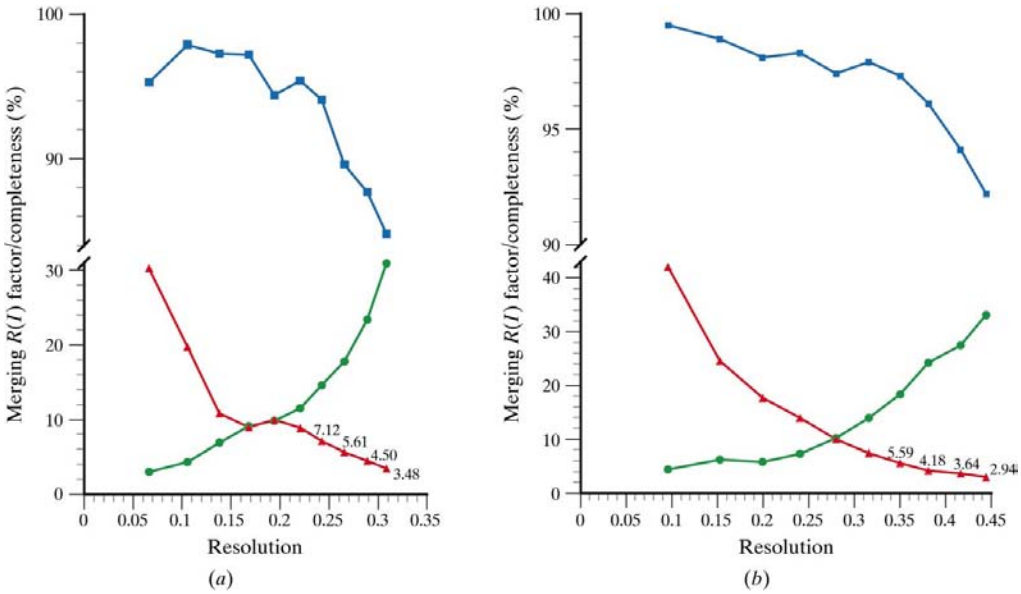

Figure 1
Completeness, $I/\sigma(I)$ and $R_{\rm merge}$ for (*a*) RT and (*b*) cryogenic data as a function of resolution. Squares indicate the completeness and triangles indicate $I/\sigma(I)$ in the resolution bins. Circles show the merging R factor $R_{\rm merge} = \sum_h \sum_i |I_{hi} - \langle I_h \rangle| / \sum_h \sum_i |\langle I_h \rangle|$, where I_{hi} is the intensity of the ith measurement of the same reflection and $\langle I_h \rangle$ is the mean observed intensity for that reflection.

conditions. Anion-exchange and affinity chromatography permitted the purification of the β-mannanase and enabled the identification of the first 42 N-terminal amino acids by Edman degradation. However, efforts to clone the respective mannanase gene have failed. Major obstacles were the asymmetric GC distribution in this thermophilic organism (approximately 70% GC content) and the existence of perhaps three enzymes of highly similar amino-acid composition (E. Tajana, unpublished results).

Despite the lack of further sequence information, we recently described the crystal structure of *T. fusca* β-mannanase at 1.5 Å resolu-

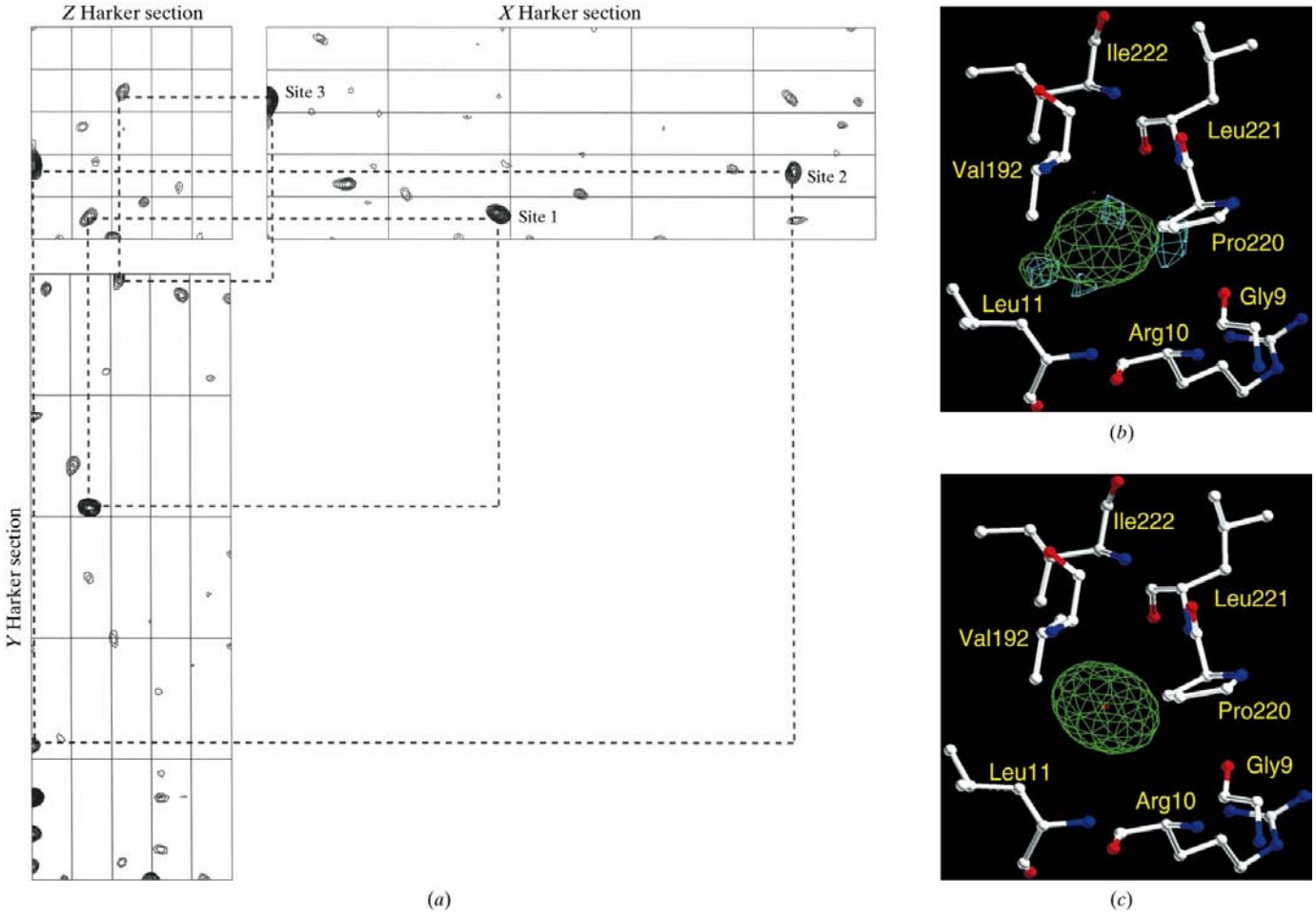

Figure 2
Xenon derivative. (*a*) Xenon Harker sections plotted for data between 8 and 2.15 Å resolution. The maps are contoured from 2σ in steps of 0.25σ. Consistent Harker peaks are connected by dashed lines. Residual maps of the strongest xenon site (*b*) before and (*c*) after anisotropic refinement.

tion (Hilge *et al.*, 1998) and provided, with the help of two complexes, a rationale for the substrate specificity in glycosyl hydrolase family 5. Here, we report the DNA sequence of *T. fusca* mannanase as obtained by a combined process of molecular biological and crystallographic methods. The experimental procedure that resulted in an electron-density map suitable for such an approach is presented in detail.

2. Materials and methods

2.1. Crystallization, data collection and processing

Three different crystal forms (forms I, II and III) were obtained under similar conditions with ammonium sulfate as the precipitant (Hilge *et al.*, 1998). They belong to space group $P2_12_12_1$ and differ mainly in the length of their *b* and *c* axes (crystal form I, $a = 43.70$, $b = 46.06$, $c = 132.51$ Å; crystal form II, $a = 46.73$, $b = 61.17$, $c = 128.53$ Å; crystal form III, $a = 46.48$, $b = 71.93$, $c = 97.45$ Å). The crystal morphologies of form I and II are rod-shaped and macroscopically indistinguishable from each other, while crystals of form III have a brick-like shape. Form III crystals were exclusively found in one single drop, together with crystals of the other forms. The solvent contents were roughly estimated to be 30, 49 and 43%, respectively, assuming an SDS–PAGE derived molecular mass of 38 kDa and one molecule per asymmetric unit. Native data from crystal form I were collected at the EMBL beamlines at HASYLAB, DESY, Hamburg to 1.8 Å at 298 K (X31, $\lambda = 0.99$ Å) and to 1.5 Å at 100 K (X11, $\lambda = 0.90$ Å). Integration and scaling of the diffraction intensities was performed with the *HKL* package (Otwinowski & Minor, 1997). The quality of the data is illustrated in Fig. 1. To summarize, the room-temperature (RT) data set was 93.4% complete and had an $R_{\rm merge}(I)$ of 7.3%, whereas the corresponding values for the cryogenic data set were 97.2 and 6.2%, respectively.

```
1                                                  10                                                  20
        GGN YTN CAY GTN AAR AAY GG
        GGG CTG CAT GTA AAG AAC GGC CGC CTG TAT GAG GCC AAC GGG CAG GAG TTC ATC ATC CGT GGC
A   T   G   L   H   V   K   N   G   R   L   Y   E   A   N   G   Q   E   F   I   I   R   G

                        30                                                  40
                        TGG TAY CCN CAR CAY CAN CA
GTC AGC CAC CCC CAC AAC TGG TAC CCC CAG CAC ACC CAG GCG TTC GCC GAC ATC AAG TCG CAC GGC GCC
V   S   H   P   H   N   W   Y   P   Q   H   T   Q   A   F   A   D   I   K   S   H   G   A

            50                                                  60
AAC ACC GTC CGG GTG GTG CTG AGC AAC GGT GTC CGG TGG AGC AAG AAC GGT CCT TCT GAC GTC GCC AAC
N   T   V   R   V   V   L   S   N   G   V   R   W   S   K   N   G   P   S   D   V   A   N

70                                                  80                                                  90
GTC ATC TCC CTG TGC AAG CAG AAC CGC CTT ATC TGC ATG CTG GAG GTG CAC GAC ACC ACC GGC TAC GGT
V   I   S   L   C   K   Q   N   R   L   I   C   M   L   E   V   H   D   T   T   G   Y   G

                            100                                                 110
GAG CAG AGC GGG GCC TCC ACG CTC GAC CAG GCC GTC GAC TAC TGG ATC GAG CTG AAG AGC GTG CTC CAG
E   Q   S   G   A   S   T   L   D   Q   A   V   D   Y   W   I   E   L   K   S   V   L   Q

                120                                                 130
GGC GAG GAG GAC TAT GTC CTC ATC AAC ATC GGC AAC GAG CCC TAC GGC AAC GAC TCC GCG ACC GTC GCC
G   E   E   D   Y   V   L   I   N   I   G   N   E   P   Y   G   N   D   S   A   T   V   A

    140                                                 150                                                 160
GCT GGG GCG TGG GAC ACC TCC GCC GCC ATC CAG CGG CTG CGC GCC GCC GGA TTC GAG CAC ACC CTC GTG
A   G   A   W   D   T   S   A   A   I   Q   R   L   R   A   A   G   F   E   H   T   L   V

                            170                                                 180
GTG GAC GCC CCC AAC TGG GGC CAG GAC TGG ACG AAC ACC ATG CGG AAC AAC GCC GAC CAG GTG TAC GCC
V   D   A   P   N   W   G   Q   D   W   T   N   T   M   R   N   N   A   D   Q   V   Y   A

                190                                                 200
AGC GAC CCC ACC GGC AAC ACC GTC TTC TCG ATC CAC ATG TAC GGC GTC TAC TCC CAG GCG TCC ACG ATC
S   D   P   T   G   N   T   V   F   S   I   H   M   Y   G   V   Y   S   Q   A   S   T   I

        210                                                 220                                                 230
ACC AGC TAC CTG GAG CAC TTC GTC AAC GCG GGC CTG CCG CTC ATC ATC GGC GAG TTC GGC CAC GAC CAC
T   S   Y   L   E   H   F   V   N   A   G   L   P   L   I   I   G   E   F   G   H   D   H

                                240                                                 250
TCC GAC GGC AAC CCC GAC GAG GAC ACG ATC ATG GCC GAG GCC GAG CGG CTC AAG CTG GGC TAC ATC GGC
S   D   G   N   P   D   E   D   T   I   M   A   E   A   E   R   L   K   L   G   Y   I   G

                        260                                                 270
TGG TCG TGG AGT GGC AAC GGC GGC GGG GTC GAG TAC CTC GAC ATG GTG TAC AAC TTC GAC GGC GAC AAC
W   S   W   S   G   N   G   G   G   V   E   Y   L   D   M   V   Y   N   F   D   G   D   N

            280                                                 290
                        ATH TTY TAY GGI CCI RAY GGI ATH
CTG AGC CCG TGG GGC GAG CGG ATA TTC TAC GGC CCC GAC GGC ATA
L   S   P   W   G   E   R   I   F   Y   G   P   D   G   I   A   S   V/T A   K   Glx A   V/T

300
I   F   G
```

Figure 3
DNA and amino-acid sequence of *T. fusca* β-mannanase as derived from Edman degradation (residues 1–42), sequencing of a PCR clone (residues 3–291) and interpretation of the electron-density map (residues 292–302). Used upstream and downstream primers are underlined, with Y = C/T, R = A/G, H = A/C/T, N = A/C/G/T and I = inosine.

2.2. Multiple isomorphous replacement including anomalous scattering

In order to solve the structure of crystal form I, the method of multiple isomorphous replacement, including anomalous scattering differences (MIRAS), was used. Trimethyllead acetate (TMLA), HgI_4, Baker's dimercurial (Baker) and $IrCl_6$ heavy-atom derivatives (herein collectively referred to as 'RT derivatives') were collected at RT to 2.4, 2.3, 3.0 and 2.5 Å resolution, respectively, whereas a xenon derivative was collected at 100 K to 2.15 Å with 25% glycerol as the cryoprotectant (Hilge *et al.*, 1998). Out of the 37 RT heavy-atom positions, only a single TMLA site was consistent in the three Harker sections. All other heavy-atom positions were exclusively located from maximum-likelihood residual maps. Compared with the crowded difference Patterson maps of the RT derivatives, the xenon Patterson map with three un-

ambiguous sites (Fig. 2*a*) was rather clean.

In order to obtain a good starting map, it was crucial that the heavy-atom refinement was performed with the program *SHARP* (de La Fortelle & Bricogne, 1997). Not only did *SHARP* help to find more weak sites utilizing the maximum-likelihood residual maps, but it also proved to be essential for simultaneous refinement of both the RT and cryogenic derivatives as well as for the inclusion of two native data sets (RT to 1.8 Å and cryogenic to 1.5 Å resolution). Initial phases were first derived from the RT data, which enabled the determination of the heavy-atom sites in the cryogenic xenon derivative. Anisotropic treatment of the strongest xenon site was important. As a consequence, the negative electron density depicted in blue (Fig. 2*b*) disappeared (Fig. 2*c*) and the isomorphous phasing power increased from 1.08 to 1.2. A comparison of the heavy-atom positions in the TMLA, Baker and $IrCl_6$ derivatives revealed that the four strongest TMLA sites were also present in the two other derivatives although with different occupancies. In contrast, the sites found for the HgI_4 and the xenon derivatives were unique.

2.3. Density modification with *wARP* and autobuilding of the polypeptide backbone

After completion of the heavy-atom parameter refinement, phases were gradually extended by solvent flipping starting from the resolution where the figure of merit dropped below 0.5 (2.44 Å) to the maximal resolution of 1.5 Å using the *SOLOMON* (Abrahams & Leslie, 1996) protocol as implemented in *SHARP*. Best results were obtained with 30% solvent content. The electron-density map was clearly interpretable and the *wARP* procedure (Perrakis *et al.*, 1997) from the *ARP*/*wARP* software suite was used to further improve the phases. Six free-atom models were built, refined for 80 cycles and thereafter averaged as described in Perrakis *et al.* (1997). The corresponding electron-density map was of excellent quality. An at that time experimental version of the autobuilding module of *ARP*/*wARP* permitted autotracing of the main-chain fragments for 292 residues and produced a guess for the amino-acid sequence. The amino-acid sequence of the likely related mannanase (SWISS-PROT code 51529) of *Streptomyces lividans* helped to manually connect the fragments and assemble them in a globular molecule. At this stage, only two loop regions (residues 230–233 and 261–266) were missing.

2.4. PCR cloning, sequencing and sequencing from the electron density

2.4.1. N-terminal sequencing. The 42 N-terminal amino acids of *T. fusca* mannanase were derived from Edman degradation and enabled the design of degenerate upstream and downstream PCR primers. The upstream primer (GGN YTN CAY GTN AAR AAY GG) encoded the amino-acid sequence GLHVKNG (Fig. 3), while the downstream primer (TGG TAY CCN CAR CAY CAN CA) coded for the amino-acid sequence WYPQHTQA (Fig. 3). 40 cycles of PCR (denaturation at 368 K for 1 min, annealing at 311 K for 1 min, extension at 345 K for 1 min) with a 5 min hot start were run using 250 ng *T. fusca* genomic DNA as template. Thereafter, 1µl of reaction mixture was taken as template for another 40 PCR cycles under the same conditions. A fragment of the expected size (∼110 base pairs) was amplified, cloned into a pUC19 vector and sequenced with an automated LI-COR 4000L sequencer using cycle sequencing in order to prevent G/C-rich related compressions.

2.4.2. Structural sequencing. The amino-acid sequence was derived as closely as possible from the available electron-density map using the program *O* (Jones *et al.*, 1991). The identification was straightforward with the exception of differentiation between Asp/Asn/Leu, Glu/Gln and Val/Thr, for which the distinction was predominantly based on their environment. This model was subjected to maximum-likelihood refinement as implemented in *REFMAC* (Murshudov *et al.*, 1997). After each refinement cycle, water molecules were automatically integrated into the model with *ARP* (Lamzin & Wilson, 1993).

2.4.3. Structure-assisted cloning and sequencing. In order to verify the sequence determined from the electron-density map after two rounds of model building and *REFMAC*/*ARP* refinement, the polypeptide chain was inspected for a stretch of eight amino acids ideally closely located to the C-terminus and with the lowest possible degree of degeneration on the DNA-base level. The amino acids Ile284–Ile291 (Fig. 4) matched these requirements and allowed the synthesis of a

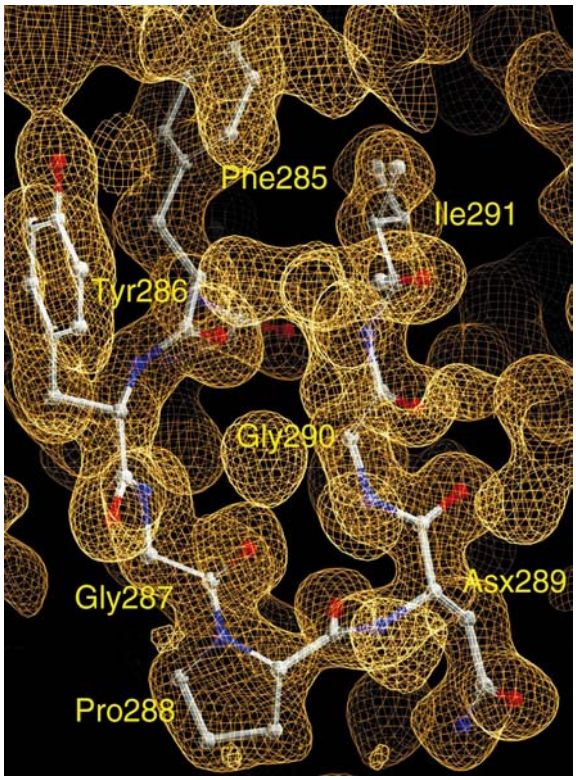

Figure 4
Electron density of the eight amino-acid stretch (Ile284-Phe285-Tyr286-Gly287-Pro288-Asx289-Gly290-Ile291) close to the C-terminus.

144-fold degenerate downstream oligonucleotide (ATH TTY TAY GGI CCI RAY GGI ATH; Fig. 3). Together with the specific upstream oligonucleotide (GGG CTG CAT GTA AAG AAC GG; Fig. 3) coding for the amino-acid sequence GLHVKNG, 2 × 40 cycles of PCR were run (conditions above). A faint band matching the expected size was cloned into the pUC19 vector and 867 base pairs corresponding to 289 amino acids (Fig. 3) were sequenced in a forward and reverse reaction. The polypeptide sequence previously assigned from the electron-density map was then checked and where necessary corrected. The DNA sequence has been deposited with the EMBL data bank (accession number AJ006227).

2.5. Anisotropic refinement

After convergence of the isotropic refinement, riding H atoms were included and individual atomic temperature factors were modelled anisotropically with *REFMAC* 4.0 (Murshudov *et al.*, 1999) using all data. Restraints used in isotropic and anisotropic refinement were 0.02 Å for the bond lengths and 0.04 Å for the bond angles. Anisotropic temperature factors were restrained according to the standard criteria in *REFMAC* and the weights were optimized to achieve a maximum drop in the free *R* factor.

For Ser65, Ser72, Glu109, Val113, Gln115, Ser134, Gln181, Ile195, Gln203, Thr206, Glu245, Lys248 and Asn271 double conformations were built, while for the two flexible loop regions (residues 230–233 after β-strand 7 and 260–262 after β-strand 8) only the higher occupied loop with a fixed occupancy of 0.7 could be modelled.

2.6. Fully automated model building

In an attempt to further prove and quantify the usefulness of automated model building as implemented in the *ARP/wARP* suite, the *warpNtrace* (Perrakis *et al.*, 1999) procedure was run starting from the *SOLOMON*-derived phases with the DNA-derived amino-acid sequence. Only default values were used for the calculations. Automatic map construction, interpretation of the map as a free-atoms model, refinement of the free-atoms model and finally five autobuilding cycles each followed by ten refinement cycles were performed without any manual intervention.

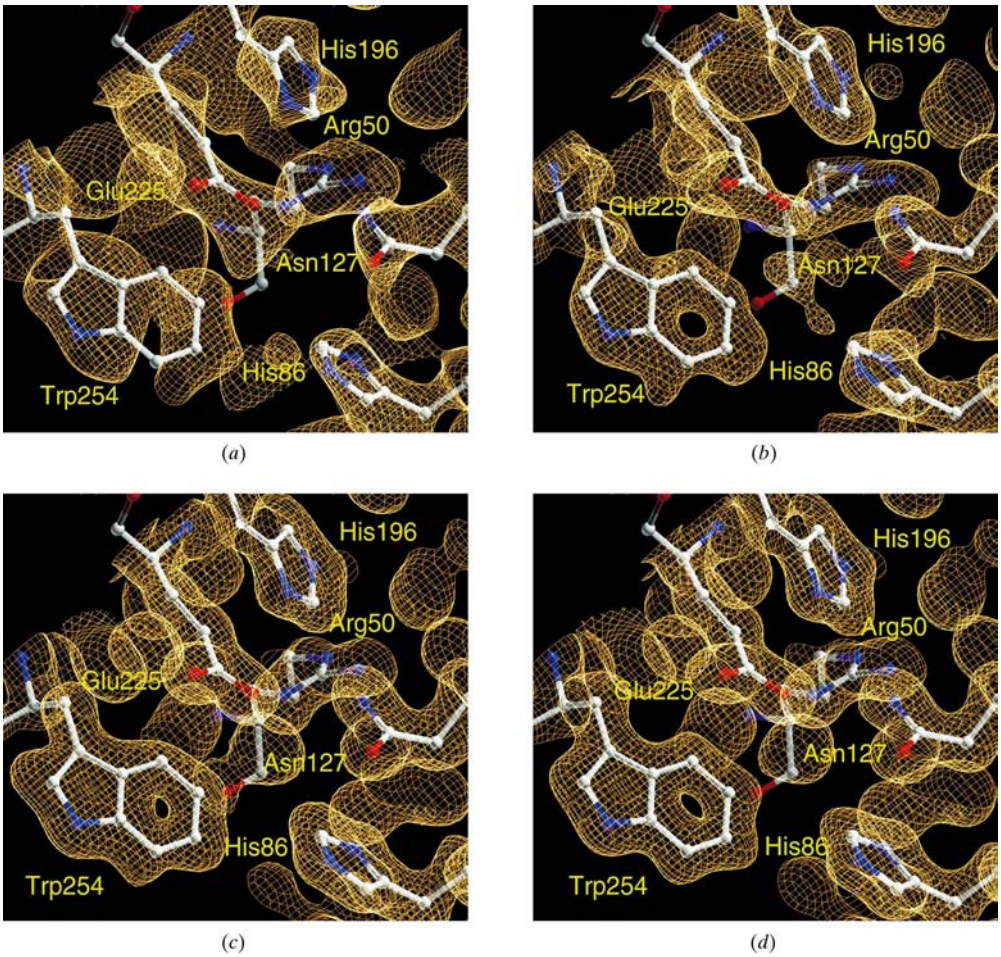

Figure 5
Progress during structure determination. (*a*) to (*d*) display the −1 subsite of the *T. fusca* mannanase active site. The amino-acid residues shown are from the final model. Electron-density maps are contoured at a level of 1σ for (*a*) heavy-atom refinement with *SHARP*, (*b*) density modification and phase extension to 1.5 Å resolution with *SOLOMON* and (*c*) phase improvement with *wARP*. (*d*) The final electron-density map of the anisotropically refined model.

3. Results and discussion

3.1. MIRAS phasing, phase improvement and extension

The crystal structure of crystal form I of *T. fusca* β-mannanase was determined by the MIRAS method. The mannanase exhibits the fold of the classical $\alpha_8\beta_8$ barrel first found in chicken triose phosphate isomerase (Banner *et al.*, 1975) and is almost spherical in shape, with dimensions of 45 × 45 × 40 Å. The progress during structure determination is illustrated in Figs. 5 and 6.

RT derivatives with TMLA, HgI_4, Baker and $IrCl_6$ could only be obtained after replacement of the precipitant ammonium sulfate by 2.0 *M* magnesium sulfate. This phenomenon has already been reported in the literature (Blake, 1968) and may be related to the competition between NH_4^+ ions and the heavy-atom compounds. Despite their very similar unit-cell parameters (deviations smaller than 0.2%), the RT

derivatives were non-isomorphous with the cryogenic xenon derivative and could initially not be used together. This could only be achieved with *SHARP*. The reason for the incompatibility most probably lies in the different solvent structure of the RT and the cryogenic crystal structure. Consequently, the additional phase information of the xenon derivative was mainly useful between 2.35 and 2.15 Å (Fig. 6). Taking advantage of the high-resolution cryogenic data, the good phases to about 2.4 Å and the properly calculated phase probability distributions in *SHARP*, solvent flattening with *SOLOMON* greatly improved the quality of the electron-density map. Furthermore, phases could be safely extended to as far as 1.5 Å resolution (Figs. 5*b* and 6). Prior to any inspection of the electron-density map, phases were improved with *wARP* (Figs. 5*c* and 6). Especially in view of the missing amino-acid sequence, the high-resolution maps obtained from *ARP*/*wARP* proved to be very helpful and substantially shortened the time spent interpreting and building the mannanase structure. In particular, building of a first model was further postponed and considerably simplified through the autobuilding of several main-chain fragments with the by this time developed *ARP*/*wARP* autobuilding modules. Finally, the likely related *S. lividans* mannanase sequence helped to connect the main-chain fragments.

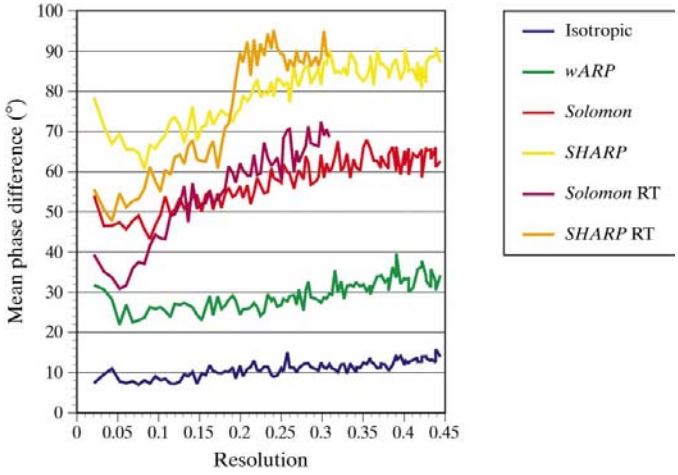

Figure 6
Mean phase differences calculated from phases between the final anisotropically refined *T. fusca* mannanase structure and phases available during the structure-solution process (see text). Phase differences between sets were calculated with *SFTOOLS* (B. Hazes, unpublished program).

3.2. Deduction of the protein and DNA sequence

Since O, N and C atoms have similar atomic form factors and the differences in bond lengths are small, distinction between these atoms is difficult even at 1.5 Å resolution. Therefore, the differentiation between Asp/Asn/Leu, Glu/Gln and Val/Thr was based mainly on hydrogen bonds and salt bridges and to a limited extent on sequence comparisons (there are eight strictly conserved amino acids in glycosyl hydrolase family 5; Wang *et al.*, 1993; Ducros *et al.*, 1995). Additionally, at this resolution individual *B* factors may help to distinguish between a valine and a threonine and, with less certainty, between N and O in Glu/Gln and Asp/Asn. Generally, *B* factors for C, N and O are balanced in correctly built side chains.

In order to obtain a correct unambiguous sequence assignment, direct sequencing of the gene is necessary. After two rounds of model building and *REFMAC*/*ARP* refinement, the free *R* factor fell to 21.5%. At this point, it was possible to identify a continuous stretch of eight amino acids at the C-terminus, allowing the determination of 95.7% of the mannanase gene sequence by PCR (§2.4). The structure-derived sequence was then corrected using the amino-acid sequence deduced from the DNA sequence. Apart from the N-terminal 42 amino acids, 171 residues (69%) of the remaining polypeptide chain were correctly identified from the electron-density map. Further refinement revealed the two loops missing from the initial model and resulted in a model containing 302 residues.

According to the MALDI-MS-derived molecular weight and comparisons with the closely related *S. lividans* sequence, an extension of about 30 amino acids at the C-terminus was expected. Indeed, a considerable amount of residual electron density was found in the region beyond Gly302, but extensive trials to build the carboxy-terminal end of the polypeptide failed. Many cellulases, xylanases and mannanases possess a cellulose-binding domain (CBD; Tomme *et al.*, 1995) either attached to the N- or C-terminus. This CBD is connected by a 20–60 amino-acid-long PT-rich linker which is often glycosylated in order to protect against proteolytic attack. Since some glycosyl hydrolases missing a CBD kept at least part of the linker, this region might be disordered owing to its inherent flexibility.

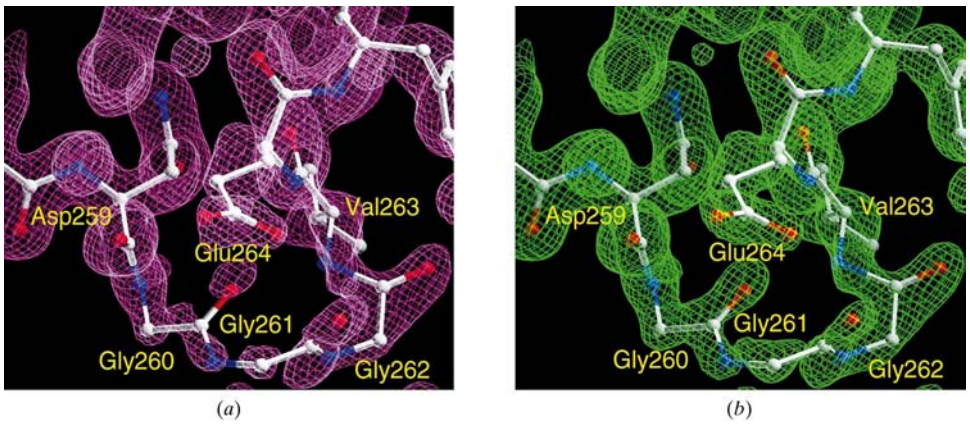

Figure 7
(*a*) Isotropic and (*b*) anisotropic refinement of individual atomic temperature factors.

The inclusion of riding H atoms and individual atomic *B*-factor refinement decreased $R_{\rm cryst}$ and $R_{\rm free}$ by 4.6 and 2.5% to the final values of 11.9 and 17.6%, respectively. The maximum and the minimum of the σ_A-weighted difference Fourier synthesis ($mF_o - DF_c$) were reduced from 0.59 and -0.40 e Å^{-3} to 0.49 and -0.38 e Å^{-3}, respectively, therefore lowering the noise considerably without worsening the model geometry. The positive effects of the anisotropic refinement are illustrated in Fig. 7, which shows the flexible loop region of residues 260–264.

A comparison of the fully refined semi-automatically derived model with the completely autobuilt model was performed using *LSQKAB* (Kabsch, 1976) and shows that the average displacement for atomic coordinates is 0.031 Å for C^{α} atoms, 0.038 Å for main-chain atoms and 0.154 Å for all atoms. The regions missing in the autobuilt model were amino acids 229–234 (disordered in the final structure), residues 254–265 (residues 260–262 disordered in the final structure) and residues 274–275. The initial autobuilding found 292 residues, 11 more than present in the final 'assessment' run. This was purely a consequence of the conservative parameters chosen in the final *ARP*/*wARP* autobuilding module in order to avoid the introduction of errors. The completely autobuilt model has no discrepancies with respect to the final model other than the missing residues. The entire procedure took around 8 h of CPU time on a Pentium III computer running Linux.

4. Conclusions

An efficient procedure to solve and refine a crystal structure as well as to clone and subsequently derive the DNA/amino-acid sequence of a protein was demonstrated by combining methods from crystallography and molecular biology. Gene cloning can be facilitated by the design of degenerate PCR primers based on the unambiguous interpretation of a high-resolution and high-quality electron-density map. The use of modern software such as *SHARP*, *SOLOMON* and *wARP* can provide an electron-density map and a model very close to the fully refined structure within a few CPU hours. Thereby, high-resolution native diffraction data are essential for the auto-building (*ARP*/*wARP*) part of the procedure as outlined in Perrakis *et al.* (1999). Even for large proteins 'structure-based DNA walking', *i.e.* identification and use of internal amino-acid sequences, could be possible.

We are very grateful to Eric de La Fortelle, Victor Lamzin and Garib Murshudov for their extensive help and enthusiasm with the programs *SHARP*, *ARP* and *REFMAC*, respectively. We would also like to thank Professor Christoph Kratky and Oliver Sauer for their help in collecting the xenon-derivative data. We particularly thank the people at the EMBL outstation at DESY in Hamburg and at the Swiss Norwegian Beamline at ESRF in Grenoble for their support. Alexander Kraev, Lei Li and Fredi Stutz are thanked for competent help in DNA sequencing. This work was funded by ETH grant 0-20-158-96 and Dr W. Kolb AG, CH-8903 Hedingen.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Waley, S. G. (1975). *Nature (London)*, **255**, 609–14.
Blake, C. C. F. (1968). *Adv. Protein Chem.* **23**, 59.
Ducros, V., Czjzek, M., Belaich, A., Gaudin, C., Fierobe, H.-P., Belaich, J.-P., Davis, G. J. & Haser, R. (1995). *Structure*, **3**, 939–949.
Hilge, M., Gloor, S. M., Rypniewski, W., Sauer, O., Heightman, T. D., Zimmermann, W., Winterhalter, K. & Piontek, K. (1998). *Structure*, **6**, 1433–1444.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**,472–494.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
McCleary, B. V. (1988). *Methods Enzymol.* **160**, 596–614.
Murshudov, G., Vagin, A. & Dodson, E. (1997). *Acta Cryst.* D**53**, 240–255.
Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–55.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.
Tomme, P., Warren, R. A. & Gilkes, N. R. (1995). *Adv. Microb. Physiol.* **37**, 1–81.
Wang, Q., Tull, D., Meinkes, A., Gilkes, N. R., Warren, R. A. J., Aebersold, R. & Withers, S. G. (1993). *J. Biol. Chem.* **268**, 14096–14102.