Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Re-refinement using reprocessed data to improve the quality of the structure: a case study involving garlic lectin

Gosu Ramachandraiah,[a] Nagasuma R. Chandra,[b] A. Surolia[a] and M. Vijayan[a]*

[a]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India, and [b]Bioinformatics Centre, Indian Institute of Science, Bangalore 560012, India

Correspondence e-mail: mv@mbu.iisc.ernet.in

Received 5 July 2001
Accepted 13 December 2001

PDB Reference: garlic lectin, 1kj1, r1kj1sf.

The structure of dimeric garlic lectin was previously determined to an effective resolution of 2.8 Å using X-ray intensity data processed by the *XDS* package and refined using *X-PLOR* [Chandra *et al.* (1999), *J. Mol. Biol.* **285**, 1157–1168]. Repeated attempts to grow better crystals with a view to improving the definition of the structure did not succeed. The available raw data were then reprocessed using *DENZO*. The structure was re-refined with both *X-PLOR* and *CNS* separately using the reprocessed data, which extended to a resolution of 2.2 Å. These two sets of refinements and the two sets using the *XDS*-processed data afforded an opportunity to compare the performance of different data-processing and refinement packages when dealing with data from weakly diffracting crystals. The best results were obtained when *CNS* was employed for refinement using data processed by *DENZO*. The quality and the resolution of the map and the definition of the structure improved substantially. In particular, the amino-acid residues at the variable locations in the sequence, and hence the isolectins, could be identified with a high degree of confidence. It could be established that the crystal asymmetric unit contains two identical heterodimers. The new refined structure also provided a better definition of other finer structural details.

1. Introduction

As part of our long-term programme on plant lectins (Banerjee *et al.*, 1996; Sankaranarayanan *et al.*, 1996; Suresh *et al.*, 1997; Prabhu *et al.*, 1998; Manoj *et al.*, 2000), we have recently reported the crystal structure of a mannose-specific agglutinin from garlic at 2.4 Å resolution (Chandra *et al.*, 1999). The structure revealed how variations in quaternary structure might be exploited by nature as a strategy to bring about differences in oligosaccharide specificity (Vijayan & Chandra, 1999). The study also indicated the presence of two different heterodimers in the asymmetric unit and of multiple mannose-binding sites in them. However, some issues still remained unresolved, largely owing to the limited resolution of the diffraction data. The most crucial of these issues is perhaps the inability to identify the correct sequence of the entire polypeptide chain and hence the correct isolectin.

Simultaneous occurrence of several isolectins is a common phenomenon in plants, the isolectins differing subtly from each other not only in their primary structures (Barre *et al.*, 1996) but also in their fine specificities for oligosaccharides. In addition, garlic lectin is known to undergo post-translational processing at several stages, leading to differences at the C-terminus and thus to polypeptide chains of varying lengths.

Multivalency and oligomerization in lectins are two important parameters through which variety in carbohydrate recognition is achieved. Bulb lectins, in particular, belong to that category of lectins where multiple carbohydrate-recognition motifs exist in the same subunit. For example, garlic lectin has three mannose-binding sites per 109-residue subunit. In addition to these established carbohydrate-binding sites, the bulb lectins are also thought to have several subsites that assist in the binding of higher mannose structures. These subsites are expected to be different not only in different species but also in two related isolectins. Differences between isolectins therefore have enormous implications for their function (Ciopraga *et al.*, 2000), thus making it important to identify the correct isolectin by identifying both the correct sequence and the total length of the polypeptide chain.

Since attempts to grow better crystals in order to collect higher resolution data were not successful, we explored the possibility of extracting the maximum information from the existing data. We have therefore reprocessed the data and refined the structure again using different methods. A detailed comparison of the four end-structures obtained using the two commonly used data-processing packages *XDS* (Kabsch, 1988) and *DENZO* (Otwinowski & Minor, 1997) combined with the two widely used structure-refinement packages *X-PLOR* (Brünger, 1992*a*) and *CNS* (Brünger *et al.*, 1998) is presented here. In addition to their immediate relevance to garlic lectin, we believe these results will be of general interest in protein crystallography in relation to extracting the best possible information from available data (Kleywegt, 2000).

2. Methods

Purification, crystallization and data collection of the mannose-specific agglutinin from garlic have been described previously (Dam *et al.*, 1998; Chandra *et al.*, 1997, 1999). X-ray diffraction data collected on a MAR imaging-plate detector have been reprocessed and the structure re-refined. The original data collection was interrupted by a power failure after recording 100 frames of 1° rotation each. A further set of 102 frames were subsequently recorded. The merging of the two data sets resulted in a comparatively high $R_{\rm merge}$, without any substantial improvement in the completeness of non-zero reflections. Therefore, only the first set was used in the published analysis. The discrepancy was considerably lower when *DENZO* was used for processing. Furthermore, it was subsequently felt that even when *XDS* is used, the improved completeness of all data adequately compensated for the higher $R_{\rm merge}$ obtained on using the entire 202 frames. Data recorded on all 202 frames and processed separately using *XDS* and the *HKL* suite of programs (*DENZO* and *SCALEPACK*) have been used in the present analysis. The structure was refined independently using *X-PLOR* and *CNS* employing separately *XDS*-processed and *DENZO*-processed data sets. The previously reported coordinates were used as the starting model (PDB code 1bwu; Berman *et al.*, 2000). Structure refinement was coupled with several iterations of manual rebuilding using *FRODO* (Jones, 1978). The parameters that were used include applying bulk-solvent and *B*-scale corrections as well as different refinement methods such as using a 'residual target' and a 'maximum-likelihood target'. Fig. 1 depicts the various methods employed. The four end-structures obtained through *XDS–XPLOR*, *XDS–CNS*, *DENZO–XPLOR* and *DENZO–CNS* are termed XX, XC, DX and DC, respectively. *SFCHECK* (Vaguine *et al.*, 1999), *PROCHECK* (Laskowski *et al.*, 1993), *WHAT-IF* (Hooft *et al.*, 1996) and the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994) were

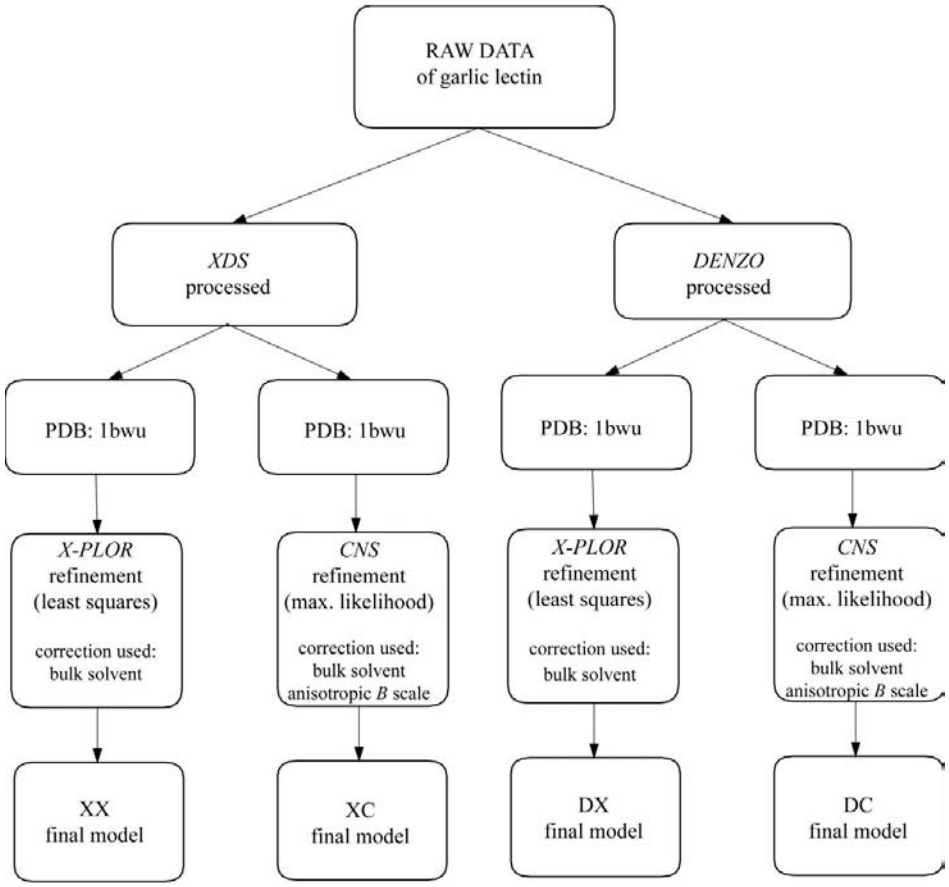

Figure 1
Schematic representation of the methods employed in obtaining the four end-models XX, XC, DX and DC.

used for evaluating the quality of the structure-factor data and the atomic models.

3. Results and discussion

3.1. Data quality

Data processed using *MAR-XDS* indicated a *C*2 lattice with unit-cell parameters $a = 202.9$, $b = 43.8$, $c = 79.2$ Å, $\beta = 112.4^\circ$. The same raw data processed using *DENZO* resulted in a *C*2 lattice of unit-cell parameters $a = 201.8$, $b = 43.5$, $c = 78.7$ Å, $\beta = 112.3^\circ$, showing a clear reduction in the unit-cell volume of 1.7%. The crystal-to-detector distance was refined in both cases. The refined values were 149.7 and 149.4 mm, respectively. The slight difference in the two values is not large enough to fully account for the difference in unit-cell parameters.

Statistics pertaining to the processing of data using the two program suites are given in Table 1. There was no data at resolution higher than 2.4 Å in the set processed by *XDS*. Even within this limit, the proportion of reflections with positive intensities was very low in the higher resolution shells. Processing using *DENZO*, however, led to a significant increase in the resolution of the data. As reported in Table 1, the data could be processed comfortably to 2.2 Å with 44.2% positive intensities in the highest resolution bin. Estimation of the background while detecting the spots in the diffraction data is established to be superior to that in *XDS*, owing to its ability to estimate the background for each spot unlike the average background for each frame as in *XDS*, and also owing to its flexible weighted profile-fitting algorithm (Otwinowski & Minor, 1997). The quality of the data as judged by $\langle I\rangle/\sigma(I)$ ratio was also considerably superior. A plot of the natural logarithm of the mean values of F_{obs} for constant values of $|h|$, $|k|$ and $|l|$ as a function of average inverse resolution, computed using *DATAMAN* (*Uppsala* suite of programs; Kleywegt & Jones, 1996), for *DENZO*-processed data is shown in Fig. 2. The plot clearly indicates a high degree of anisotropy in the data. The plot using *XDS*-processed data also indicated a similar level of anisotropy. The importance of correcting for anisotropy either during data processing or during structure

Table 1
Data-collection statistics.

Values in parentheses refer to the highest resolution shell.

	XDS	*DENZO*	
Space group	*C*2	*C*2	
Unit-cell parameters			
a (Å)	202.9	201.8	
b (Å)	43.8	43.5	
c (Å)	79.2	78.7	
β (°)	112.4	112.3	
Z	8	8	
Unit-cell volume (Å^3)	651210	640010	
Solvent content (%)	61.9	59.6	
Data processing (Å)	2.4	2.2	2.4
No. of observations	103216	102203	92792
No. of unique reflections	25078 (5197)	29430 (1791)	24529 (4611)
No. of reflections with $I = 0$	12307 (4520)	2001 (300)	1245 (589)
Completeness of data (%)	97.6 (96.1)	90.3 (55.8)	97.4 (92.9)
Completeness when reflections with $I = 0$ are omitted	47.9 (13.2)	80.7 (44.2)	88.7 (77.2)
Multiplicity	4.1	3.5	3.7
Merging *R* for all reflections (%)	9.4 (54.2)	10.4 (43.6)	10.7 (39.6)
Average $I/\sigma(I)$	12.9 (0.9)	14.4 (2.3)	16.8 (3.3)

† For the *DENZO*-processed data, statistics for data cutoff at 2.4 Å are also given in the subcolumn to enable direct comparison with the *XDS*-processed data (2.60–2.40 and 2.28–2.20 Å).

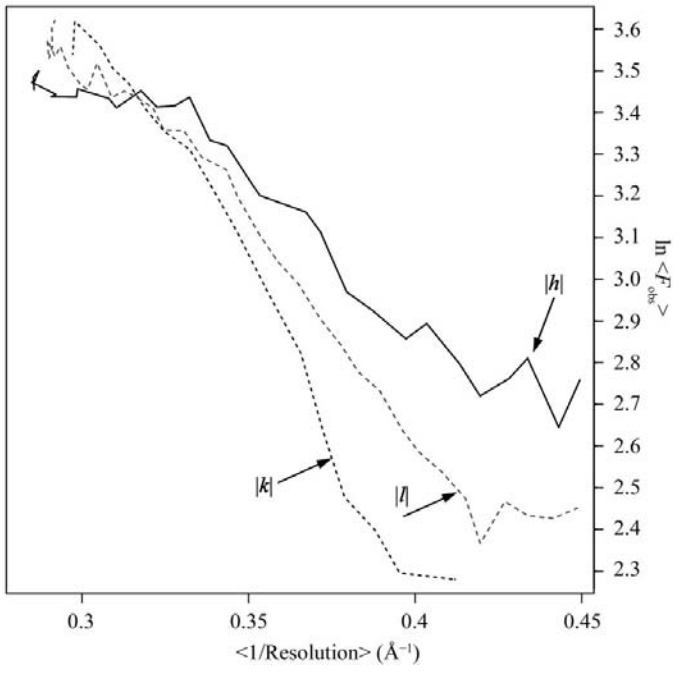

Figure 2
Plot of the natural logarithm of the mean values of F_{obs} corresponding to constant values of $|h|$, $|k|$ and $|l|$ with respect to the average value of 1/resolution. As the maximum $|h|$ is as high as 91, only every third value has been used in the plot corresponding to this index. The figure was prepared using *DATAMAN* (Kleywegt & Jones, 1996) from the *Uppsala* suite of programs.

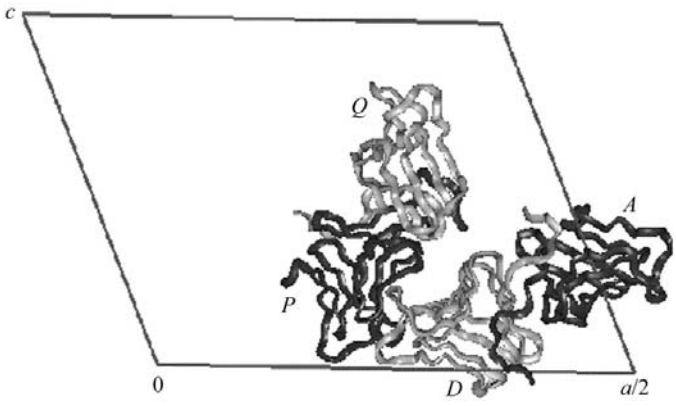

Figure 3
Crystal structure of garlic lectin. The two dimers *AD* and *PQ* in the asymmetric unit are shown in ribbon representation viewed down the *b* axis.

Table 2
Refinement parameters for the four models.

	XX model	XC model	DX model	DC model
Maximum resolution (Å)	2.4	2.4	2.2	2.2
No. of reflections with $F > 0$	12771	12771	27423	27423
Final R factor (%)	25.6	21.8	27.6	21.1
$R_{\rm free}$ (%)	34.2	29.5	34.5	25.1
$\langle u \rangle$ error in coordinates by Luzzati plot† (Å)	0.41	0.40	0.37	0.32
RMS deviations from ideal				
Bond length (Å)	0.118	0.010	0.009	0.007
Bond angles (°)	1.9	1.8	1.9	1.6
Dihedral angles (°)	25.8	26.2	27.0	26.9
Improper angles (°)	1.0	0.9	1.0	0.7
No. of protein atoms	3359	3359	3388	3402
No. of sugar atoms	168	168	168	168
No. of solvent atoms	165	151	176	175

† Luzzati (1952).

refinement to minimize errors in the final model has been described previously (Murshudov *et al.*, 1998).

3.2. Comparison of the four models

The asymmetric unit contains two dimers (Fig. 3). Subunits *A* and *D* form one dimer, while subunits *P* and *Q* form the other. The RMS deviations in C^{α} atoms for various combinations of subunit pairs from all four models ranged from 0.3 to 0.7 Å for *AP* or *DQ* pairs, but ranged between 0.8 and 1.0 Å for *AD* or *PQ* pairs. This is consistent with the observation that subunits *A* and *P* (or *D* and *Q*) have identical sequences but subunits *A* and *D* (or *P* and *Q*) have several differences between them (see also §3.2.1).

Table 2 summarizes the refinement statistics for the four models. It is clear that the *DENZO*-processed data refined with *CNS* is the best of the four different models as judged by crystallographic *R* factors and $R_{\rm free}$ values (Brünger, 1992*b*). An increase in the number of reflections with positive intensities in the *DENZO*-processed data has automatically led to an increase in the observations-to-parameters ratio, a metric that influences the quality of the refined structure. The important features that are available with *CNS* but not with *X-PLOR* are (*a*) use of a maximum-likelihood target function and (*b*) the inclusion of anisotropic scaling. In order to assess the comparative importance of these features, independent refinements were carried out using *DENZO*-processed data and *CNS* with (i) isotropic scaling (IS) and least-squares minimization (LSM), (ii) isotropic scaling and maximum-likelihood minimization (MLM), (iii) anisotropic scaling (AS) and least-squares minimization and (iv) anisotropic scaling and maximum-likelihood minimization. The variation of *R* and $R_{\rm free}$ as a function of resolution was similar in all four calculations. Anisotropic scaling led to a reduction in *R* and $R_{\rm free}$ of more than 3% irrespective of the minimization function used, while LSM and MLM led to comparable *R* and $R_{\rm free}$, whichever scaling function was used. Thus, it appeared that it was primarily the inclusion of AS that led to better refinement using *CNS*.

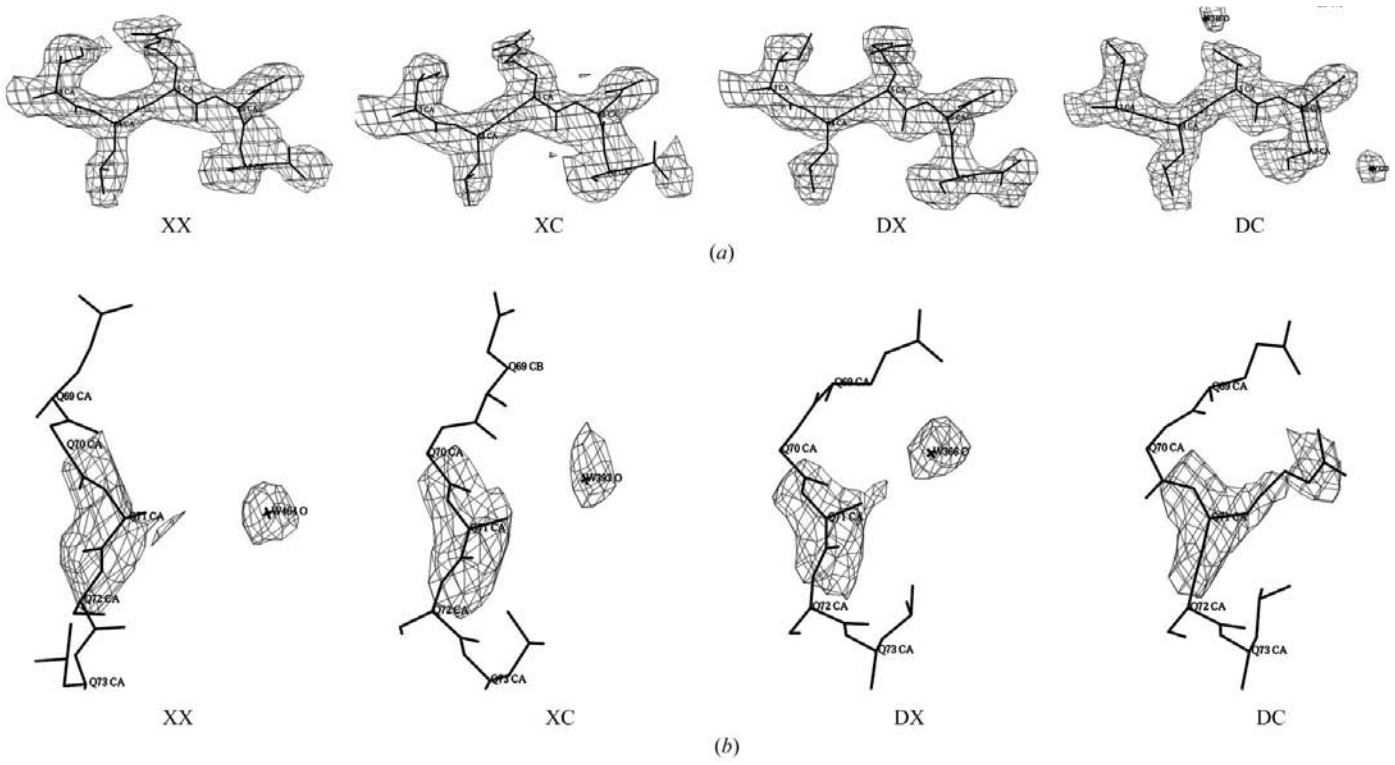

Figure 4
Examples to indicate the improvement in clarity in the electron-density maps of the four models. The two regions shown are (*a*) residues 3–7 in subunit *A* and (*b*) residue 71 in subunit *Q*. See text for details.

3.2.1. Structural features. Fig. 4 illustrates an example of the improvement in clarity in the electron-density map of the DC model as described here. Residue 5 was inferred to be an arginine in subunits *A* and *P* and to be a threonine in *D* and *Q* in the XX model. Comparison of the sequence of garlic lectin reported in the literature and determined by us previously by sequencing the genomic DNA segments (Chandra *et al.*, 1999) indicated that residue 5 could be an arginine, a threonine or a methionine. This was not too surprising, given that the molecule was a heterodimer and that the sample could also contain a mixture of isolectins. From our previous study, we identified an arginine in one subunit and a threonine in the other three subunits, which together with some other residue differences implied that not only was the molecule a heterodimer but also that the two dimers in the asymmetric unit were not identical. Given that several isolectins coexist in the plant, this issue could not be resolved any further despite the crystal structure. The DC model, however, shows clear electron density at this position in all four subunits, corresponding to a threonine in one subunit but a methionine in the other subunit, in both the dimers. 11 such disparities have been resolved in the DC model, thus resulting in the identification of two identical heterodimers (*AD* and *PQ*) in the asymmetric unit. The two subunits *A* and *D* in a heterodimer differ from each other in ten of 109 residues (Fig. 5), namely, residues 3, 5, 7, 23, 39, 42, 43, 46, 48 and 106. In the XX model, no clear electron density was observed for the four C-terminal residues in the subunits *A* and *P*, which implied that the heterodimers *AD* or *PQ* could contain subunits of slightly different polypeptide lengths. The DC model, however, clearly shows that all subunits have 109 residues and are therefore of equal length.

The C-terminal residues Tyr108 and Lys109 are close to one of the mannose-binding sites in all subunits, but it is interesting to note that only in one subunit (per dimer) are they oriented appropriately for interactions with the mannose hydroxyl groups. The loop regions have been better defined in the DC model, leading to rebuilding of regions 17–21, 35–37, 66–68 and 98–100. Some additional interactions have also become clearly discernible from the DC model. The most notable are the three additional salt bridges (Fig. 6*a*) between residues Asp17 and His22, between residues Asp67 and Arg71 in each subunit and between residues Asp35 of one subunit and Lys109 of the adjacent subunit. These residues are found in the loop regions between the sheets and could contribute significantly to stabilizing the loops in that particular orientation, which in turn may be important for higher oligosaccharide binding.

Improvement in the electron-density maps has also led to more precise identification of water molecules. Three tryptophans are situated at the centre of the subunit, close to the threefold axis, forming a hydrophobic core. A network of four water molecules within hydrogen-bonding distance of NE1 of the tryptophans is seen in every subunit (Fig. 6*b*). It is interesting to note that fluorescence studies carried out to study the folding and unfolding of this protein by Surolia and coworkers reveal a significant red shift (λ_{344}) in the unfolded state compared with λ_{328} in the folded state (Bachhawat *et al.*, 2001). This observation can be explained by the presence of water molecules interacting with the tryptophans, although the exact structural role of these waters is still not clear.

Bulb lectins have an exclusive specificity for mannose and mannose-rich oligosaccharides, a property which enables them to mediate several biological processes. Three mannose-binding sites were identified in every subunit of the garlic lectin structure. While the electron density was very clear for one of them and was considered to be the primary binding site, the other two sugars refined with partial occupancy. In the DC model, however, electron density for all three sugars was clear and could be refined with full occupancy. While the interactions of the sugars with the protein remain the same, a few additional water molecules have been identified in the DC model.

Apart from the three mannose sugars per subunit, an additional mannose molecule per dimer was observed close to the binding site on sheet III. Since the site on sheet III is considered to be the biologically more relevant site (Wright & Hester, 1996), we had proposed that the additional site in our previous model could perhaps be a biologically important subsite. Mannose sugars are bound on the surface of each sheet in a well

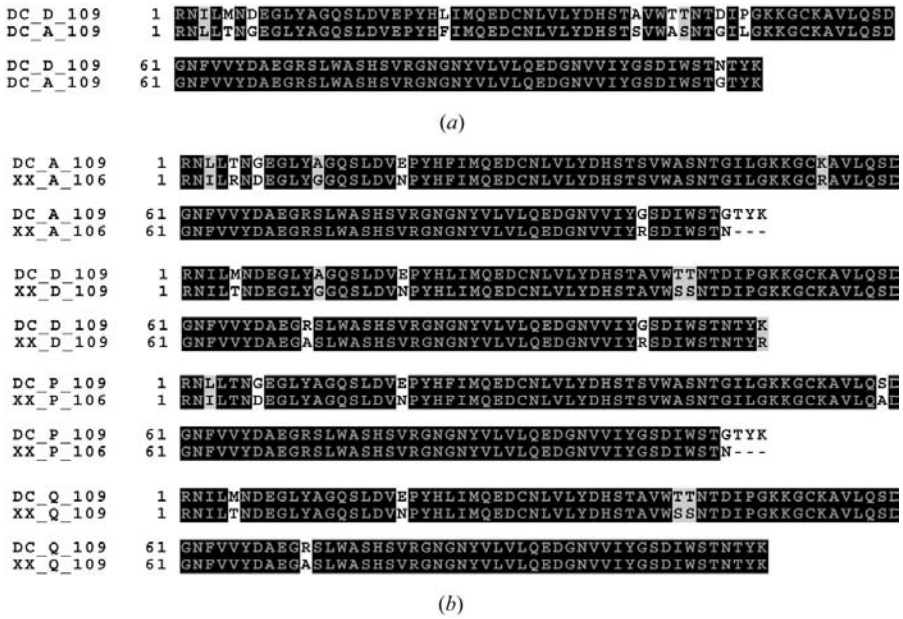

Figure 5
(*a*) Sequence alignment of the two subunits *A* and *D* in a heterodimer as deciphered with the aid of the DC model. (*b*) Comparison of the sequences deduced in all the four subunits between the XX and DC models. The alignment shown is based on the structure. The figure was prepared using *BOXSHADE* (Hofmann & Baron, 2001).

defined groove characterized by both a sequence signature and a structural motif (Ramachandraiah & Chandra, 2000). The additional subsite, despite not having the features of these signatures, contained a few key residues, namely Asp35, Tyr21 and Val40, which have been shown to be crucial for mannose binding. Subsequent to our study, the crystal structure of daffodil lectin, a closely related protein reported by Sauerborn *et al.* (1999), also exhibited a similar phenomenon albeit at a different location. In the DC model, however, electron density for this sugar, though present, has not improved any further, if not worsened compared with that in the XX model. Residue 93, a glycine which is in close proximity to the sugar, if identified as an arginine, can account for majority of the electron density at this additional site. All sequencing studies so far have indicated this residue to invariably be a glycine and we have therefore had little reason to build an arginine here. However, caution needs to be exercised before any biological interpretations are made based on this density alone. It may be that a higher resolution structure or a complex with mannose-rich oligosaccharides is required to throw more light on this issue.

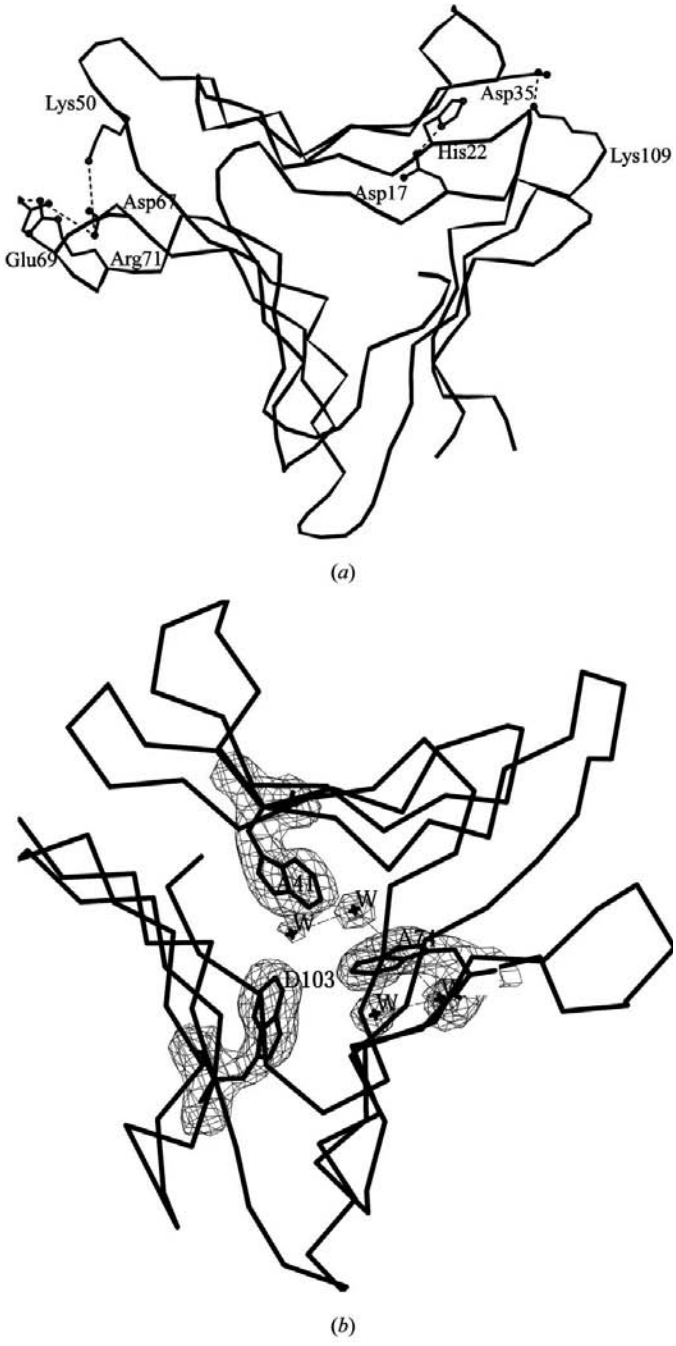

Figure 6
(*a*) Salt bridges in a subunit of garlic lectin. The C^{α} trace of subunit *D* is shown in grey and the segment from subunit *A* is shown in black. Side chains involved in salt-bridge formation are shown as ball-and-stick models. (*b*) Network of water molecules interacting with the tryptophans present close to the internal threefold axis of the molecule.

4. Conclusions

The results presented here indicate that the completeness, resolution and quality of the data can be significantly improved by careful data processing. The flexibility available through the set of programs in *DENZO* is particularly useful here. While there is no general rule, structure-refinement strategies can be made more effective if the data is analyzed for particular properties such as anisotropy so that appropriate corrections can be applied, as illustrated by the DC model of garlic lectin.

Improvement in the data quality and hence the electron-density map has resulted in the resolution of several sequence ambiguities in garlic lectin and thus in the identification of the correct isolectin. The presence of two identical heterodimers in the asymmetric unit has also been clearly determined. Rebuilding of the loops has led to the identification of three salt bridges that may be important for their stability. Overall, the reliability and the accuracy of the final model has improved substantially. The model also provides a plausible explanation for the results of the fluorescence studies on the folding and unfolding of the protein.

We thank Professor T. P. Singh for suggesting the re-refinement of the structure using reprocessed data. The computations and model building were carried out at the Super Computer Education and Research Centre at the Indian Institute of Science and the Bioinformatics Centre and the Graphics facility supported by the Department of Biotechnology. GR is a CSIR senior research fellow. The work is supported by the Department of Science and Technology.

References

Bachhawat, K., Dam, T. K., Kapoor, M. & Surolia, A. (2001). *Biochemistry*, **40**, 7291–7300.
Banerjee, R., Das, K., Ravishankar, R., Suguna, K., Surolia, A. & Vijayan, M. (1996). *J. Mol. Biol.* **259**, 281–296.
Barre, A., Van Damme, E. J. M., Peumans, W. J. & Rouge, P. (1996). *Plant Physiol.* **112**, 1531–1540.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Brünger, A. T. (1992*a*). *X-PLOR. Version 3.1. A System for X-ray Crystallography and NMR.* Yale University, Connecticut, USA.
Brünger, A. T. (1992*b*). *Nature (London)*, **355**, 472–475.
Brünger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, D., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Chandra, N. R., Dam, T. K., Surolia, A. & Vijayan, M. (1997). *Acta Cryst.* D**53**, 787–788.
Chandra, N. R., Ramachandraiah, G., Bachhawat, K., Dam, T. K., Surolia, A. & Vijayan, M. (1999). *J. Mol. Biol.* **285**, 1157–1168.
Ciopraga, J., Angstrom, J., Bergstrom, J., Larsson, T., Karlsson, N., Motas, C., Gozia, O. & Teneberg, S. (2000). *J. Biochem.* **128**, 855–867.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dam, T. K., Bachhawat, K., Geetha Rani, P. & Surolia, A. (1998). *J. Biol. Chem.* **273**, 5528–5535.
Hofmann, K. & Baron, M. D. (2001). *BOXSHADE: Printouts from Aligned Protein or DNA Sequences*. http://bioweb.pasteur.fr/seqanal/interfaces/boxshade.html.
Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.
Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.
Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
Manoj, N., Srinivas, V. R., Surolia, A., Vijayan, M. & Suguna, K. (2000). *J. Mol. Biol.* **302**, 1129–1137.
Murshudov, G. N., Davies, G. J., Isupov, M., Krzywda, S. & Dodson, E. J. (1998). *CCP4 Newsl.* **35**, 37–42
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Prabhu, M. M., Sankaranarayanan, R., Puri, K. D., Sharma, V., Surolia, A., Vijayan, M. & Suguna, K. (1998). *J. Mol. Biol.* **276**, 787–796.
Ramachandraiah, G. & Chandra, N. R. (2000). *Proteins Struct. Funct. Genet.* **39**, 358–364.
Sankaranarayanan, R., Sekar, K., Banerjee, R., Sharma, V., Surolia, A. & Vijayan, M. (1996). *Nature Struct. Biol.* **3**, 596–603.
Sauerborn, M. K., Wright, L. M., Reynolds, C. D., Grossmann, J. G. & Rizkallah, P. J. (1999). *J. Mol. Biol.* **290**, 185–199.
Suresh, A. S., Geetha Rani, P., Pratap, J. V., Sankaranarayanan, R., Surolia, A. & Vijayan, M. (1997). *Acta Cryst.* D**53**, 469–471.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Vijayan, M. & Chandra, N. (1999). *Curr. Opin. Struct. Biol.* **9**, 707–714.
Wright, C. S. & Hester, G. (1996). *Structure*, **4**, 1339–1352.