Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

MUMBO: a protein-design approach to crystallographic model building and refinement

Martin T. Stiebritz and Yves A. Muller*

Lehrstuhl für Biotechnik, Institut für Biologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Henkestrasse 91, D-91052 Erlangen, Germany

Correspondence e-mail: ymuller@biologie.uni-erlangen.de

Received 25 November 2005
Accepted 12 April 2006

In recent years, significant progress has been achieved in automation of the crystal structure-determination process. However, the final part of this process, namely the refinement of the atomic model, is still tedious for biological macromolecules because, especially at lower resolution, it requires extensive manual intervention. Here, it is shown that computer algorithms widely used in protein-design approaches can substantially simplify this process, helping to identify the correct orientation of the side chains during refinement. This approach was implemented into the computer program *MUMBO*. As in many protein-design programs, side-chain rotamer diversity is generated using rotamer libraries. The selection of the best combination of side-chain orientations is based on either the dead-end elimination (DEE) theorem or a Metropolis Monte Carlo approach and on a detailed atomic scoring function that describes the molecular interactions between the rotamers. We show that this scoring function can be easily extended and complemented through the introduction of an X-ray pseudo-energy calculated from the electron density present at the position of the rotamer. This extension is fully compatible with present protein-design algorithms and it is shown for a number of test cases that using this approach model refinement is simplified and convergence occurs faster.

1. Introduction

In the present heyday of numerous structural genomics initiatives, considerable effort is being invested in automating the individual steps of the crystal structure-determination process. Whereas some of the steps such as protein production, crystallization and data collection greatly profit from the introduction of robotics, other steps such as phasing and model building have benefitted from the development of sophisticated computer programs such as *ARP*/*wARP* (Morris *et al.*, 2004; Perrakis *et al.*, 1999), *RESOLVE* (Terwilliger, 2000) and *MAID* (Levitt, 2001). These programs allow very rapid derivation of atomic models of the structures to be solved. However, especially at resolution lower than 2.5 Å these models are often incomplete and manual intervention is required. Automated refinement procedures are typically alternated with manual rebuilding and correction of the atomic model until agreement between the model coordinates and the experimental diffraction data is achieved (Drenth, 1994; Tronrud, 2004). This process can be very time-consuming and typically proves to be more tedious the lower the resolution of the diffraction data set.

© 2006 International Union of Crystallography
Printed in Denmark – all rights reserved

Table 1
MUMBO program steps.

Program step	Description
INIT	Initialization of the atomic structure and generation of side-chain diversity
MC	Elimination of those rotamers that make unfavourable interactions with the constant part of the macromolecule (De Maeyer *et al.*, 2000)
DEE	Elimination of rotamers using the dead-end elimination theorem of zeroth order (Desmet *et al.*, 1992)
GOLD	Elimination of rotamers using the dead-end elimination theorem of first order (Goldstein, 1994)
SPLIT	Elimination of rotamers using conformational splitting (Pierce *et al.*, 2000)
DOUB	Identification of rotamer dead-ending pairs followed by another round of Goldstein elimination (Gordon *et al.*, 2003)
MONT	Searching for a solution of the side-chain placement problem using a Metropolis Monte Carlo approach
BRUTE	Explicit calculation of the overall energies of all remaining possible combinations of rotamers and identification of the rotamer combination with the lowest energy
ANA	Detailed analysis of the side-chain rotamer interaction energies

In parallel to this progress in crystallography, significant achievements have been obtained in a different field of structural biology, namely in the area of *de novo* protein design. In particular, the use of side-chain packing and selection algorithms has proved to be very successful for the computational *de novo* design of proteins with novel or considerably altered properties (Dahiyat & Mayo, 1997; Dwyer *et al.*, 2004; Kuhlman *et al.*, 2003; Looger *et al.*, 2003).

A typical challenge in protein design is to identify a novel primary sequence that, when produced as a polypeptide, will fold into a given tertiary structure. In principle, the sequence that best fits a given backbone architecture can be identified through the systematic variation of the amino acids and their side-chain orientations at every position of the protein backbone and by identifying the combination of amino acids that displays the most favourable energetic interactions. Although this problem cannot be solved directly because of the infinite number of potential side-chain orientations, several numerical approaches have proved successful in overcoming this problem (Canutescu *et al.*, 2003; Desmet *et al.*, 1992; Simons *et al.*, 1999). These rely on the use of rotamer libraries, *i.e.* on the discrete sampling of the conformational space an amino-acid side chain can occupy. This is warranted because soon after the first crystal structures of proteins had been solved, it was noticed that the orientations of the side chains cluster into defined rotameric states (see, for example, Ponder & Richards, 1987). To date, various rotamer libraries differing in size and complexity have been compiled and are commonly used in protein-design projects (Dunbrack & Cohen, 1997; Lovell *et al.*, 2000). However, a further reduction of the combinatorial complexity is needed. Desmet and coworkers showed that one possibility is to use the dead-end elimination theorem (DEE; Desmet *et al.*, 1992). Without having to calculate explicitly the energy content of every single combination, the DEE theorem identifies and eliminates those rotamers that cannot be part of the global minimum-energy conformation (GMEC). Further improvements and extensions of this basic procedure by various research groups culminated in the first *de novo* designed structure, namely of a zinc-finger motif published in 1997 (Dahiyat & Mayo, 1997). Other approaches used in protein design rely on stochastic methods such as Monte Carlo algorithms, which cannot guarantee the identification of the GMEC but usually find configurations close to it.

Because the process of fitting residue side chains into electron-density maps during the crystallographic model-building step is highly comparable to identifying the optimal side-chain conformations in computational protein design, we have implemented the dead-end elimination theorem, several extensions thereof and a Metropolis Monte Carlo algorithm into the computer program *MUMBO* (Table 1). The program reads in electron-density maps and atomic coordinates and determines the GMEC or a configuration close to the GMEC of amino acids and side-chain orientations while taking into account classical energetic contributions such as van der Waals interaction, electrostatic interaction, atom solvation and hydrogen-bonding energy as well as an X-ray pseudo-energy derived from the electron density present at the positions of the atoms of a given rotamer. We show that when applying this approach to a number of test cases, atomic model building is simplified and the crystallographic refinement converges rapidly.

2. Theoretical background

2.1. Generation of side-chain diversity

In order to find the best configuration of side-chain orientations (that is, the combination representing the GMEC), three distinct topics have to be addressed. Firstly, all possible orientations that a side chain can display have to be generated. Secondly, an atomic force field has to be defined that allows accurate description of the energetic interactions between the different side chains. Thirdly a selection process has to be implemented that allows identification of the combination of the side-chain orientations with the lowest overall energy and which therefore should correspond to that present in the crystal structure.

As for other protein-design programs, *MUMBO* relies on the use of rotamer libraries and a user can choose between a backbone φ and ψ angle-independent (Lovell *et al.*, 2000) and a backbone conformation-dependent rotamer library (Dunbrack & Cohen, 1997). In addition, a fine-step option allows further expansion of the conformational space. In this case, user-specified angle increments are added to and subtracted from each dihedral angle to generate additional rotamers. In case hydrogen-bonding energies are to be evaluated between side chains (see below), polar H atoms are placed automatically into the model and the number of rotamers is expanded if the position of the H atom to be added is ambiguous. Thus, three different hydrogen positions are generated for the hydroxyl hydrogen of the amino acids serine and threonine and two for tyrosine.

2.2. Side-chain interaction energies

A semi-empirical knowledge-based force field is used to calculate the overall energy of the system. So far, the following contributions are considered: van der Waals attraction and repulsion energy, electrostatic interaction, hydrogen-bonding energy, side-chain solvation energy and a pseudo-energy derived from the abundance with which a given rotamer is observed in high-resolution crystal structures. These rotamer probabilities are an integral part of every rotamer library (Dunbrack & Cohen, 1997; Lovell *et al.*, 2000). In order to use the force field for crystallographic purposes, an X-ray pseudo-energy calculated from the electron density that is present at a given rotamer position was added (see below). The overall expression for the force field is given in (1) and includes weights (w) to adjust the contributions from the different energy types,

$$E_{\rm Total} = w_{\rm vdW}E_{\rm vdW} + w_{\rm Elec}E_{\rm Elec} + w_{\rm Rotprob}E_{\rm Rotprob} + w_{\rm Hbond}E_{\rm Hbond} + w_{\rm Solv}E_{\rm Solv} + w_{\rm Xray}E_{\rm Xray}. \quad (1)$$

In (1), $E_{\rm vdW}$ denotes the van der Waals energy, $E_{\rm Elec}$ the electrostatic energy, $E_{\rm Rotprob}$ a pseudo-energy derived from the rotamer probability, $E_{\rm Hbond}$ the hydrogen-bonding energy, $E_{\rm Solv}$ the solvation energy and $E_{\rm Xray}$ the X-ray pseudo-energy.

The van der Waals energy is calculated using the standard Lennard–Jones (12, 6) potential. Atom properties such as van der Waals radii and equilibrium energies are assigned using the atom-type libraries from the CHARMM19 force field (Brooks *et al.*, 1983; Neria *et al.*, 1996). These are commonly used in standard crystallographic refinement programs such as, for example, *CNS* (Brünger *et al.*, 1998). Because the interaction energies are calculated from discrete rotamers and it is therefore not possible to escape from high repulsion energies through small adjustments in the orientation of the side chains, the van der Waals repulsion energies can be artificially softened either by scaling down the atomic radii or using a softened repulsion potential (Pokala & Handel, 2005).

Different modes exist for the calculation of the electrostatic interactions. Besides using the standard Coulomb term with a user-supplied dielectric constant, it is possible to make the dielectric constant distance-dependent. Furthermore, a switch or a shift function can be applied to ensure that the electrostatic interaction between two charges becomes zero if they are further apart than a user-defined threshold. Again, the CHARMM19 force-field parameters are used to assign full and partial atom charges.

Solvation free energies are estimated according to the solvation model developed by Lazaridis & Karplus (1999). This model allows the pairwise decomposition of the solvation energies, which is a prerequisite for application of the dead-end elimination theorem (see below). Hydrogen-bonding energies can be calculated using two different models. The first corresponds to an empirical orientation-dependent hydrogen-bonding potential developed by Kortemme *et al.* (2003). In this approach, hydrogen-bonding energies are calculated *via* a linear combination of various angle- and distance-dependent tabulated values derived from high-resolution protein crystal structures. Alternatively, hydrogen-bonding energies can be calculated explicitly from donor-atom and acceptor-atom geometries. In this case, hydrogen-bonding energies are calculated based on the donor–acceptor distance and three additional angular constraints that account for the rotational degrees of freedom along the hydrogen bond, the chemical nature of the donor and acceptor atoms and the atoms to which these are connected (Gohlke *et al.*, 2004).

Finally, an X-ray pseudo-energy is calculated as the negative sum of the electron-density values present at the positions of the atoms of a given rotamer. Electron densities are read in as regular $\sigma_{\rm A}$-weighted $2F_{\rm o} - F_{\rm c}$ electron-density maps (Read, 1986) and atom electron densities $\rho(x_i, y_i, z_i)$ are calculated *via* linear interpolation from the density at the surrounding grid points (Rossmann *et al.*, 1992). The expression for the crystallographic pseudo-energy is

$$E_{\rm Xray} = -\sum_{{\rm Atoms},\,i} \rho(x_i, y_i, z_i). \quad (2)$$

2.3. Identification of the global minimum-energy conformation (GMEC)

The main challenge for any side-chain packing algorithm is to cope with the huge number of possible rotamer combinations. For a small protein with 100 residues and with an average of five distinct side-chain orientations per residue, the number of combinations is as large as 5^{100}.

Alogrithms based on the DEE theorem or on stochastic methods such as Metropolis Monte Carlo optimization provide an adequate means to find a solution for this extraordinary combinatorial problem. An important precondition for their application is the assumption that the overall protein energy can be decomposed into the energy of the backbone, the energy of the individual residues and the pairwise interaction energy between different residues,

$$E_{\rm Total} = E_{\rm Template} + \sum_i E(i) + \sum_{j>i} E(i, j). \quad (3)$$

Whereas $E_{\rm Template}$, the energy of the backbone, is constant for all possible combinations and therefore can be omitted during the calculations, $E(i)$ denotes the energy of the amino acid at position i, *i.e.* its self energy (chemical bonds) and its interaction with the backbone, and $E(i, j)$ represents the pairwise interaction energy between the amino acids at positions i and j. The dead-end elimination theorem now provides a criterion to decide whether a certain rotamer i_r at side-chain position i can safely be eliminated from further considerations as it cannot be part of the GMEC. This is the case if the best possible interaction energy (min term in equation 4) of this rotamer with the residues at all other positions j_s is higher than the worst possible interaction energy (max term in equation 4) of an alternate rotamer i_t,

$$E(I_r) + \sum_{j\neq i} \min_s E(i_r, j_s) > E(i_t) + \sum_{j\neq i} \max_s E(i_t, j_s). \quad (4)$$

Several improvements of this basic criterion have been published (Goldstein, 1994; Voigt *et al.*, 2001; Gordon *et al.*,

2003) and we have implemented a number of them in the program *MUMBO* (Table 1). However, it has to be noted that this procedure will not necessarily lead to a unique solution. Therefore, a brute-force approach follows the dead-end elimination steps in *MUMBO* if the remaining combinations can be evaluated in a reasonable amount of time. In this case, the energies are calculated explicitly in order to identify the combination that corresponds to the GMEC.

Despite the fact that the multiple variants of the dead-end elimination theorem were optimized for increasing elimination efficiency, the set of remaining rotamers is sometimes still too large to be explored explicitly by a brute-force approach. To overcome this problem, we implemented a stochastic algorithm to arrive at a single solution instead. The procedure consists of a Monte Carlo optimization in combination with the Metropolis criterion (Metropolis *et al.*, 1953) to avoid trapping in local energy minima. Initially, a random configuration is built. The rotameric state at one position is then changed randomly. If the energy of this new configuration is lower than the previous one, the change is accepted. Otherwise, the new configuration is accepted, if the Metropolis criterion is fulfilled, *e.g.* if

$$\exp\left(-\frac{E - E_{\mathrm{Prev}}}{kT}\right) > R, \qquad (5)$$

where E denotes the energy of the system after the change and E_{Prev} the previous energy. R is a random number in the interval $[0, 1)$ and k and T are the Boltzmann constant and the absolute temperature, respectively.

The process of randomly changing the configuration and deciding whether the change is accepted or not is repeated for a predefined number of cycles as specified in the input file (usually of the order of 10^6). The configuration with the lowest energy identified during the optimization is stored and represents the solution of the algorithm. Although it is not guaranteed that such a stochastic procedure will find the GMEC, we found that the Monte Carlo algorithm is very efficient for the side-chain packing problem, especially if it is applied after the dead-end elimination process, which reduces the search space tremendously and removes high-energy configurations. Hence, the probability increases that the Monte Carlo approach will identify a low-energy solution or even the GMEC.

The program *MUMBO* is written in ANSI Fortran 95 and is partially parallelized with OpenMP. Presently, side-chain rotamers from up to 300 residues can be optimized simultaneously on a standard dual-processor desktop PC with reasonable running time (up to several hours). The program is freely available from the authors upon request.

3. Materials and methods

3.1. Test structures

For the repacking tests and the crystallographic rebuilding and refinement calculations, three example structures from the Protein Data Bank (PDB; Berman *et al.*, 2000) were chosen for which coordinates and structure factors have been deposited, namely the data set 1thw of the sweet protein thaumatin consisting of 207 residues (1.75 Å resolution; Ko *et al.*, 1994), the data set 2hft of the extracellular domain of the human tissue factor (211 residues, 1.69 Å; Muller *et al.*, 1996) and the data set 1dpx of hen egg-white lysozyme (128 residues, 1.65 Å; Weiss *et al.*, 2000). As a fourth test case, we chose the crystal structure of the Src homology 3 (SH3) domain from the protein tyrosine kinase Lck (Lck-SH3 domain; Koga *et al.*, 1986) that we recently solved at 1.3 Å and refined to a crystallographic R_{work} of 11.5% and R_{free} of 13.8% (unpublished results). The domain consists of 62 residues and contains a total of 240 side-chain atoms.

3.2. Reproducing the packing of side chains in refined crystal structures

For the side-chain repacking studies, *MUMBO* was compared with the programs *ROSETTA* (Simons *et al.*, 1999) and *SCWRL*3.0 (Canutescu *et al.*, 2003). All *MUMBO* calculations were performed using the backbone-dependent rotamer library of Dunbrack & Cohen (1997) and exploring different weighting schemes for the energy terms in the force field. To run *ROSETTA*, the corresponding web interface of *ROSETTA DESIGN* was used (http://rosettadesign.med.unc.edu). Resfiles were generated by choosing the option NATAA for all residues of the proteins to be repacked in order to perform only side-chain packing and not protein-design calculations. To run the *SCWRL*3.0 calculations, a downloaded command-line version with default settings was used.

3.3. Crystallographic rebuilding and refinement

For the crystallographic rebuilding and refinement calculations with *MUMBO* and other crystallographic packages that were selected with which to compare the performance of *MUMBO* (see below), we first generated polyalanine models from the refined crystal structures used as test cases. In an attempt to remove model bias, the polyalanine models were subjected to several rounds of refinement either using simulated annealing with *CNS* (Brünger *et al.*, 1998) or positional refinement with *REFMAC*5 (Murshudov *et al.*, 1997) within resolution limits identical to those used subsequently during the test calculations. The resulting models were used throughout all calculations reported here to provide starting atom positions, phases and σ_{A}-weighted $2F_{\mathrm{o}} - F_{\mathrm{c}}$ electron-density maps (Read, 1986). The crystallographic R factors of these models are reported as starting R_{free} and R_{work} in the tables.

To test the crystallographic model-building properties of *MUMBO*, a shell script was written that cycles five times between the automated model building with *MUMBO* and the automated crystallographic refinement with *REFMAC*5 (Murshudov *et al.*, 1997) from the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994). Progress during the refinement was followed by monitoring R_{work}, R_{free} and the root-mean-square (r.m.s.) deviation between the atom positions of the current model and the fully

Table 2
R.m.s. and side-chain χ dihedral angle deviations between the calculated models and the final refined crystal structure of the examples Lck-SH3 domain, thaumatin, human tissue factor and hen egg-white lysozyme.

Reported is the percentage of correctly predicted χ_1 and overall χ_i side-chain dihedral angles within 20° accuracy.

	MUMBO	*ROSETTA* (Simons *et al.*, 1999)	*SCWRL*3.0 (Canutescu *et al.*, 2003)
Lck-SH3 domain			
R.m.s.d. (Å)	0.60	0.87	0.80
$\Delta\chi_1 < 20°$ (%)	92.2	84.3	88.2
All $\Delta\chi_i < 20°$ (%)	75.9	69.6	73.2
Thaumatin			
R.m.s.d. (Å)	0.86	1.00	1.11
$\Delta\chi_1 < 20°$ (%)	85.0	84.4	79.6
All $\Delta\chi_i < 20°$ (%)	65.8	67.3	61.3
Human tissue factor			
R.m.s.d. (Å)	0.88	0.87	0.96
$\Delta\chi_1 < 20°$ (%)	82.1	82.1	79.0
All $\Delta\chi_i < 20°$ (%)	69.7	71.0	66.1
Hen egg-white lysozyme			
R.m.s.d. (Å)	1.02	1.13	1.23
$\Delta\chi_1 < 20°$ (%)	88.5	83.7	87.5
All $\Delta\chi_i < 20°$ (%)	72.0	67.0	68.4

refined reference crystal structure. Within *REFMAC*5, 45 refinement cycles were calculated during each program call.

The same script was also used with *COOT* (Emsley & Cowtan, 2004). In this case, the model-building step with *MUMBO* was replaced by the 'mutate & autofit' function in *COOT*. Here, the model was again refined with *REFMAC*5 after each model-building step and model building and refinement were repeated five times. *ARP*/*wARP* (Morris *et al.*, 2004; Perrakis *et al.*, 1999) and *SOLVE*/*RESOLVE* (Terwilliger, 2000) were run independently from the above shell script, but starting from the same polyalanine models and model phases. *ARP*/*wARP* was started from the *CCP*4 interface (Collaborative Computational Project, Number 4, 1994) selecting the option 'automated model building starting from an existing model' and default parameters of the software. For the calculations with *SOLVE*/*RESOLVE*, *RESOLVE* was used in the iterative model-building mode, again applying default settings.

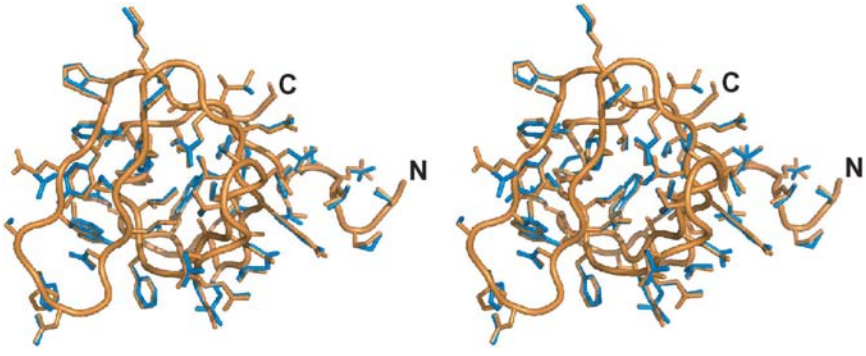


Figure 1
Stereo representation of the Lck-SH3 domain. Superposition of the reference 1.3 Å crystal structure (in blue) with the model (in orange) in which the side chains have been rebuilt with *MUMBO*. The overall r.m.s. deviation between the models is 0.6 Å.

In order to be able to compare directly the results obtained by the different program packages, we subjected all resulting models to an ultimate round of 50 cycles of refinement with *REFMAC*5 using identical geometric and crystallographic weights. This was performed to ensure that the deviations in bond lengths and angles as well as the variations in the thermal displacement factors were similar in all the models. In case of *ARP*/*wARP*, any solvent molecules automatically built by the program were removed prior to the final refinement round since the other program packages did not automatically incorporate any water molecules.

4. Results

4.1. Validation of the repacking algorithm and of the scoring function

In order to test whether the side-chain packing and selection algorithms implemented in *MUMBO* perform as expected, we attempted to reproduce the native side-chain packing of four different proteins ranging from 62 to 211 amino acids. As can be seen in Table 2, when starting solely from polyalanine models derived from the crystal structures and using the backbone-dependent rotamer library of Dunbrack & Cohen (1997), *MUMBO* is able to rebuild the side-chain orientations quite accurately in all four test cases (Table 2, Fig. 1).

The r.m.s. deviations between the repacked models and the crystal structures range between 0.6 and 1.0 Å. The accuracy of the results is also visible from the deviation of the predicted side-chain dihedral angles from those observed in the refined crystal structures. The percentage of side-chain dihedral angles predicted with 20° accuracy is generally greater than 80% for χ_1 and around 70% for all side-chain dihedral angles. *MUMBO* did very well in selecting the correct rotamers in the protein core and discrepancies are predominantly located on the protein surface (Fig. 1). This is as expected, because the energetic requirements on the surface are less restrictive compared with the core positions, so that different rotamers will display similar energies. Furthermore, *MUMBO* does not take into consideration crystal-packing contacts because the program has no knowledge of the side-chain interactions on the protein surface that are responsible for the intermolecular contacts in the crystal. The results obtained with *MUMBO* compare favourably with those obtained with established programs for side-chain placement such as *ROSETTA* (Simons *et al.*, 1999) and *SCRWL*3.0 (Canutescu *et al.*, 2003) (Table 2).

4.2. Automated model building and crystallographic refinement with *MUMBO*

Having now shown that the algorithms implemented in *MUMBO* can solve the side-chain placement problem efficiently, we investigated the use of *MUMBO* for

Table 3
Thaumatin model building with *MUMBO* using either a detailed force field ($E_{total} = \sum_i E_i$) or a force field consisting of E_{Xray} only.

			$\sum_i E_i$			E_{Xray} only		
Resolution (Å)	Starting R_{free} (%)	Starting R_{work} (%)	R_{free} (%)	R_{work} (%)	R.m.s.d. (Å)	R_{free} (%)	R_{work} (%)	R.m.s.d. (Å)
1.8	35.8	34.3	26.1	24.2	0.71	28.3	25.7	0.90
2.0	36.6	34.4	26.5	23.3	0.66	29.4	26.3	1.19
2.3	38.3	34.4	25.6	21.5	0.72	28.2	23.5	1.00
2.5	38.2	34.0	24.4	20.0	0.69	24.0	19.0	0.95
2.8	37.9	32.5	24.0	19.0	0.78	26.6	21.4	1.07

crystallographic purposes using four different test cases. In particular, we wanted to investigate how the results are influenced by the resolution of the data sets. Therefore, in separate runs, the resolution in *REFMAC*5 was restricted to the highest resolution of the particular data set and to the values 2.0, 2.3, 2.5 and 2.8 Å, respectively. The resulting models were then compared with the high-resolution crystal structures retrieved from the Protein Data Bank.

A typical time-course of an automated model building calculation with *MUMBO* is depicted in Fig. 2 for the protein thaumatin. Using data to a resolution of 1.8 Å, an accurate model is obtained rapidly. After five cycles of *MUMBO*/*REFMAC*5, R_{free} falls from 35.8% (polyalanine model) to 26.1% (Fig. 2*a*). After only one cycle, an r.m.s. deviation of 0.79 Å is obtained, which falls further to 0.71 Å after five cycles of model building (Fig. 2*c*). The results in Fig. 2 and

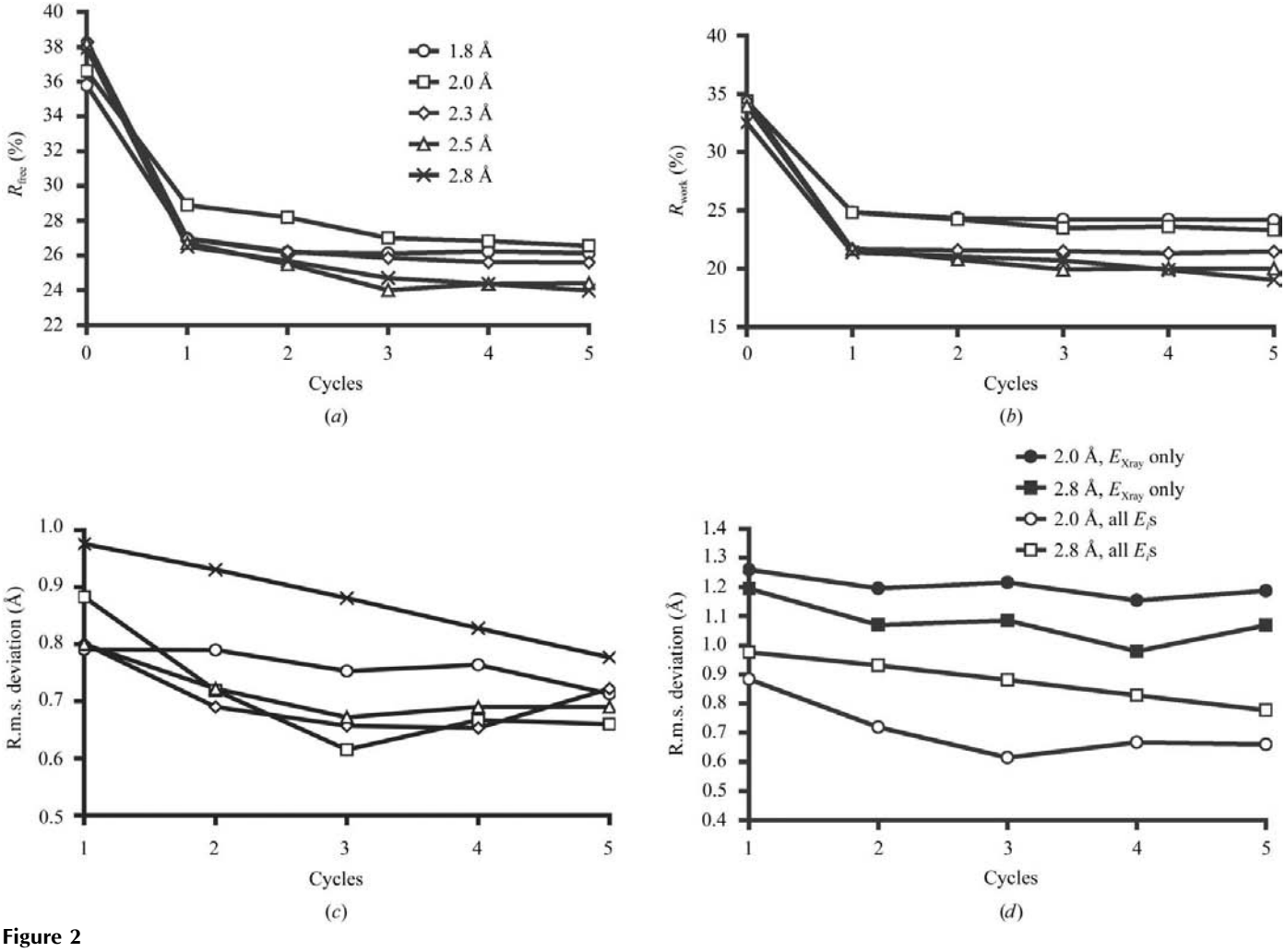
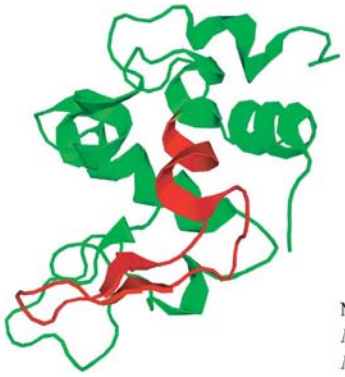


Figure 2
Automated model building of the sweet protein thaumatin with *MUMBO*. The calculations with *MUMBO* were performed at different resolution cutoffs, namely the resolution of the native data set (1.8 Å) and 2.0, 2.3, 2.5 and 2.8 Å. Five cycles of alternating between model building and refinement were performed in total for each calculation. (*a*) Progression of R_{free} over the course of the model-building calculations, (*b*) progression of R_{work} and (*c*) progression of the r.m.s. deviation between the models and the crystal structure. (*d*) Influence of the force field on the accuracy of the generated models. Depicted is the progression of the r.m.s. deviation from the crystal structure for calculations with all energy terms or with only the X-ray pseudo-energy considered in the force field at the resolutions 2.0 and 2.8 Å.



(a)

```
-2  VCAAKFESNFNTQATNRNTDGSTDY
-1  CAAKFESNFNTQATNRNTDGSTDYG
 0  AAKFESNFNTQATNRNTDGSTDYGI
 1  AKFESNFNTQATNRNTDGSTDYGIL
 2  KFESNFNTQATNRNTDGSTDYGILQ
```

(b)

```
Native sequence      31-AAKFESNFNTQATNRNTDGSTDYGI-55
MUMBO all Ei's       AAfFnSfFfTQqTNnNTnGSTsYGI
MUMBO EXray only     kkFFnSNFNqQqrrRrrnGdyyYdq
```

(c)

Figure 3
Identifying the correct sequence frame with *MUMBO*. (*a*) Cartoon representation of hen egg-white lysozyme. The fragment selected for the calculations is shown in red and includes amino acids 31–55. (*b*) Alternative sequence frames used during the calculations. The original frame has the number 0 and four alternative frames have been obtained by shifting the sequence window one and two amino acids in the direction of the N- and C-termini. (*c*) The original (native) sequence and the solutions obtained by *MUMBO* using all energy terms (E_is) and considering the X-ray pseudo-energy ($E_{\rm Xray}$) only are shown. Correctly predicted residues are depicted in uppercase bold letters and were 17 out of 25 for all energy terms and nine out of 25 residues for the X-ray pseudo-energy only.

Table 4
Summary of automated model building for three additional test cases.

Resolution (Å)	Starting $R_{\rm free}$ (%)	Starting $R_{\rm work}$ (%)	$R_{\rm free}$ (%)	$R_{\rm work}$ (%)	R.m.s.d. (Å)
Lck-SH3 domain					
1.5	43.8	41.9	33.9	32.8	0.68
2.0	42.6	41.2	32.2	30.2	0.68
2.3	42.4	41.6	31.0	29.2	0.56
2.5	45.4	40.6	31.6	28.0	0.65
2.8	45.5	38.6	33.5	29.2	0.90
Human tissue factor					
1.7	42.8	41.4	32.7	30.6	0.75
2.0	43.3	41.1	33.2	30.3	0.82
2.3	44.0	41.0	33.2	29.7	0.82
2.5	45.7	40.9	33.3	29.0	0.94
2.8	47.1	39.9	35.2	27.9	1.02
Hen egg-white lysozyme					
1.5	42.0	40.0	31.9	28.9	0.75
2.0	41.4	39.5	32.5	27.6	0.85
2.3	48.5	38.1	36.7	27.8	0.91
2.5	48.5	38.1	34.9	27.6	0.87
2.8	50.0	36.3	44.0†	30.0†	1.46†

† Failure to converge.

Table 3 show further that *MUMBO* is able to build thaumatin not only at high resolution but also if only data to low resolution are included. Even at 2.8 Å resolution, $R_{\rm free}$ falls readily, namely from 37.9 to 24.0%. The final r.m.s. deviation is 0.78 Å in this case.

The results obtained for the additional three test cases are summarized in Table 4 and confirm those already observed for thaumatin. In all cases *MUMBO* is able to readily generate complete atomic models. The mean decrease in $R_{\rm free}$ is about 11% for the Lck-SH3 domain, human tissue factor and lysozyme. The best models differ by about 0.6–0.8 Å from the fully refined crystal structures.

4.3. Importance of the force field for model-building accuracy

In order to investigate any benefits resulting from the use of a detailed force field (1), the results obtained above for the protein thaumatin were compared with calculations in which only $E_{\rm Xray}$ was used to identify the best rotamer configuration. In the latter case, the rotamers are selected based solely on the electron density present at the atom positions and this is very similar to the way rotamers are fitted into the electron density during manual model building with *O* (Jones *et al.*, 1991) and *COOT* (Emsley & Cowtan, 2004) or automatically in *ARP*/*wARP*.

As can be seen from Table 3 and Fig. 2(*d*), the accuracy of the final model decreases in the case where only $E_{\rm Xray}$ is considered to guide rotamer selection. This is particularly true when considering the r.m.s. deviations between the final models and the refined thaumatin crystal structure; these are of the order of 0.2–0.3 Å lower if a more detailed force field is used instead. However, it should be noted that the effect is smaller than one might expect because especially at low resolutions the added information content provided by the detailed force field should help to resolve ambiguities present in the electron-density maps. It is possible that in our test case, even at lower resolution, the electron densities are still significantly biased towards the refined crystal structure of thaumatin because the shifts introduced during the initial refinement of the polyalanine model might have been too small to completely remove any model bias. Nonetheless, the test calculations show that a more detailed force field improves the convergence of the refinement and the quality of the final models (Table 3).

4.4. Identifying the correct sequence registration during model building

A problem that often arises at lower resolutions in regions where the electron density is poor is how to correctly register the amino-acid sequence once the path of the backbone has been traced. This is particularly true in exposed loop regions that are often associated with high thermal displacement factors. Because protein-design algorithms such as those implemented in *MUMBO* are able to consider different amino-acid types simultaneously for a given amino-acid position in the backbone, we wondered whether the program would be able to find the correct amino-acid sequence frame for a part of a protein for which alternative sequence frames are considered. In order to test the performance of *MUMBO*, a fragment of 25 amino acids in hen egg-white lysozyme was chosen which includes a part of an α-helix, an antiparallel β-sheet and a loop region (residues 31–55; Fig. 3*a*). Residues

Table 5
Completeness (%) of the models generated by *ARP/wARP* and *SOLVE/RESOLVE*.

	Resolution (Å)				
	1.5	2.0	2.3	2.5	2.8
ARP/wARP	98	98	98	98	
SOLVE/RESOLVE	98	90	82	67	77

in this fragment were truncated to alanine and the resulting model refined at 2.5 Å resolution with *REFMAC*5 to generate a starting model and electron-density map. Different sequence frames were generated by shifting a 25-amino-acid window one or two residues in the direction of the N- and C-termini so that in principle five different amino acids have to be considered in parallel for each position in the protein fragment during the calculations (Fig. 3*b*). Automated model building with *MUMBO* was then performed at 2.5 Å resolution as previously described.

The results are summarized in Fig. 3(*c*) and show that *MUMBO* is able to identify the correct amino-acid type for 17 of the 25 positions, so that any ambiguities regarding the correct sequence registration are clearly resolved. Repeating the same calculation with only $E_{\rm Xray}$ to guide amino-acid type and rotamer selection leads to the correct identification of only nine amino acids. Furthermore, $R_{\rm free}$ of the resulting model is 4% higher in this case (37.8% *versus* 33.8%). This example again emphasizes the added benefits of using a detailed force field rather than relying only on the electron density present at a given atom position to identify the amino-acid type and its side-chain orientation.

4.5. Comparing *MUMBO* with other crystallographic software packages

For additional validation, the results obtained with *MUMBO* were compared with those from established X-ray crystallography software packages such as *ARP/wARP* (Perrakis *et al.*, 1999), *SOLVE/RESOLVE* (Terwilliger, 2000) and *COOT* (Emsley & Cowtan, 2004). It is obvious that the scope of *ARP/wARP* and *RESOLVE* goes well beyond the capabilities of *MUMBO* since, in contrast to the latter, these programs are able to trace and rebuild the entire main chain of a protein. *ARP/wARP* even includes solvent molecules into the model so that fairly complete atomic structures are obtained. Nevertheless, these programs also offer the possibility of building models from polyalanine backbones and it is this feature that we aimed to compare between the various programs. The interactive crystallographic model-building program *COOT* allows the automated fitting of rotamers into the electron density and includes an algorithm to avoid obvious steric clashes.

As a first test case, the Lck-SH3 domain was chosen for comparative model building. We started again from polyalanine models and performed the calculations at 1.5, 2.0, 2.3, 2.5 and 2.8 Å resolution (Fig. 4; Table 5). A first obvious result is that because *ARP/wARP* and *RESOLVE* rebuild parts of the backbone, they tend to remove less well defined residues from the model during the course of the calculations. While at high resolution only a few residues are removed, the number of missing residues can become substantial at lower resolutions and as a result the programs then fail to converge (Fig. 4, Tables 5 and 6). At resolutions 1.5 and 2.0 Å, highly comparable results are obtained for Lck-SH3 with all four programs. *ARP/wARP* and *MUMBO* achieve almost identical $R_{\rm free}$ and r.m.s.d. values, with slightly higher values for *RESOLVE* and *COOT* (Fig. 4 and Table 6). At a resolution of 2.5 Å or higher,

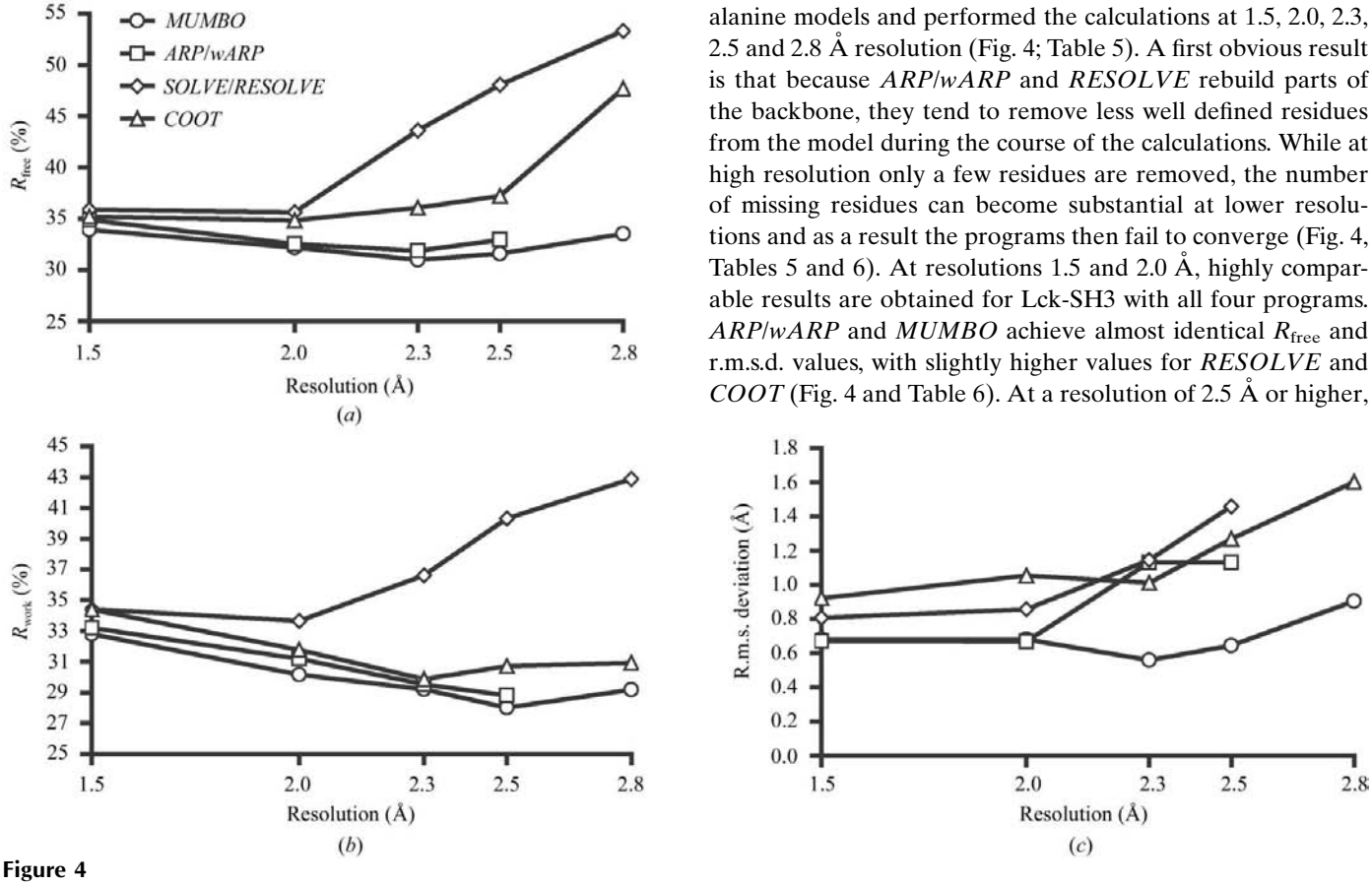


Figure 4
Comparison between *MUMBO*, *ARP/wARP*, *SOLVE/RESOLVE* and *COOT* for the model building of the Lck-SH3 domain at different resolutions. (*a*) $R_{\rm free}$ values for the models obtained by the different programs, (*b*) $R_{\rm work}$ values and (*c*) r.m.s. deviations from the refined crystal structure.

Table 6
Comparison between *MUMBO, ARP/wARP*, *SOLVE/RESOLVE* and *COOT*.

	2.0 Å					2.5 Å				
	Starting R_{free} (%)	Starting R_{work} (%)	R_{free} (%)	R_{work} (%)	R.m.s.d. (Å)	Starting R_{free} (%)	Starting R_{work} (%)	R_{free} (%)	R_{work} (%)	R.m.s.d. (Å)
Sweet protein thaumatin										
MUMBO	36.6	34.4	26.5	23.3	0.66	38.2	34.0	24.4	20.0	0.69
ARP/wARP	36.6	34.4	26.6	24.2	0.80	38.2	34.0	—	—	—
SOLVE/RESOLVE	36.6	34.4	39.7	35.5	94%†	38.2	34.0	38.4	33.3	73%†
COOT	36.6	34.4	27.5	24.8	1.16	38.2	34.0	28.3	22.8	1.15
Lck-SH3 domain										
MUMBO	42.6	41.2	32.2	30.2	0.68	45.6	40.6	31.6	28.0	0.65
ARP/wARP	42.6	41.2	32.5	31.2	0.67	45.6	40.6	32.9	28.8	1.13
SOLVE/RESOLVE	42.6	41.2	35.6	33.7	90%†	45.6	40.6	48.0	40.3	67%†
COOT	42.6	41.2	34.9	31.8	1.06	45.6	40.6	37.2	30.7	1.27
Human tissue factor										
MUMBO	43.3	41.1	32.4	29.2	0.80	45.7	40.9	32.4	27.9	1.02
ARP/wARP	43.3	41.1	31.6	28.4	99%†	45.7	40.9	47.8	40.6	54%†
SOLVE/RESOLVE	43.3	41.1	34.4	31.9	81%†	45.7	40.9	43.2	38.2	72%†
COOT	43.3	41.1	31.0	27.1	0.72	45.7	40.9	33.3	27.2	1.08
Hen egg-white lysozyme										
MUMBO	41.4	39.5	32.3	28.8	0.87	48.5	38.1	43.0	28.3	0.79
ARP/wARP	41.4	39.5	44.6	38.4	99%†	48.5	38.1	61.8	51.6	0%†
SOLVE/RESOLVE	41.4	39.5	37.5	32.4	91%†	48.5	38.1	54.5	42.3	65%†
COOT	41.4	39.5	34.3	29.8	1.23	48.5	38.1	45.5	29.8	1.47

† In these cases, only incomplete models have been built and the percentage of built residues is given instead of the r.m.s.d. value.

however, only *MUMBO* is able to generate atomic models that display low r.m.s.d.s when compared with the fully refined crystal structure. Almost identical results are also obtained for the three additional test cases (Table 6). In most cases, *MUMBO* performs best based on R_{free}, R_{work} and r.m.s.d. values and this is especially true for calculations performed at lower resolution.

5. Discussion

We have shown that the protein-design algorithms implemented in *MUMBO* can resolve the side-chain positioning problem in general and are able to reproduce for a number of test cases the orientations of the side chains observed in a crystal structure when starting solely from the correct backbone trace. The accuracy of *MUMBO* in correctly predicting the side-chain orientations slightly outperforms that of *ROSETTA* (Simons *et al.*, 1999) and *SCWRL*3.0 (Canutescu *et al.*, 2003). One observation was that the calculations were less accurate when hydrogen-bonding energies were considered explicitly. This is an unexpected result because the introduction of sophisticated energy terms describing hydrogen bonding had significantly improved the accuracy of protein-design calculations in the past (Kortemme *et al.*, 2004).

This study shows that the general side-chain packing approach can be easily extended towards crystallographic applications through the introduction of an X-ray pseudo-energy. Starting from phases obtained from an initial refinement of a polyalanine model, *MUMBO* is able to very quickly identify the correct side-chain orientations, not only at high resolution but also in the case where only data to low resolution are available. This capability must be attributed to the additional energy terms covered by the force field, because the accuracy of the models generated by *MUMBO* is significantly better when a detailed force field is used instead of only the X-ray pseudo-energy. It appears that the reduction in the number of observed data at low resolution and therefore the loss of information regarding the spatial positioning of the side chains can be compensated by only allowing physically reasonable orientations. There is certainly room for improvements in *MUMBO* regarding the X-ray pseudo-energy. The way the electron density of a given rotamer presently is converted into an energy follows a very rudimentary formalism and more sophisticated approaches such as the calculation of real-space density-correlation factors (Jones *et al.*, 1991) might provide a more accurate estimation of the X-ray pseudo-energy. Nevertheless, the examples presented here show that it is straightforward to treat electron density as an additional energy term in the framework of the dead-end elimination theorem and the Metropolis Monte Carlo approach. In general, the results obtained with *MUMBO* compare quite well with those obtained by other crystallographic programs. At low resolution, *MUMBO* outperforms the other programs in the test cases presented here.

The use of protein-design approaches during the crystallographic model-building process generates a number of novel opportunities. A common procedure in protein crystallography is to model regions with poor electron density as polyalanine or polyglycine fragments first and to only incorporate side chains in these regions once the backbone has been traced correctly. *MUMBO* can perform this automatically since protein-design algorithms implicitly allow different types of residues to be evaluated simultaneously at a specific position. If the rotamer of the correct residue does not pick up any electron density because of inaccuracies in the positioning of the main chain or if unfavorable van der Waals clashes occur, it would be easy for *MUMBO* to substitute these residues for alanine or glycine instead until in subse-

quent runs the backbone is positioned correctly and the correct sequence is then energetically favoured. We showed in a test case that alternative sequences can already be used successfully to probe alternative sequence registrations and to identify the correct sequence frame.

Protein-design algorithms extensively probe the energetic interactions of a residue with the surrounding residues. In this sense, these algorithms mimic approaches typically at the heart of structure-validation programs such as, for example, the 3D–1D profile method (Bowie *et al.*, 1991) or those used in threading algorithms (Jones *et al.*, 1992) for homology modelling. *MUMBO* was deliberately conceived as a very versatile program because it allows for the user to weight the different energetic contributions individually. Thus, if the X-ray energy is turned off, *MUMBO* can be used for validation purposes and a detailed analysis of the atomic interactions in the protein will be generated.

The aim of the work reported here was not to present the ultimate computer algorithm for automated model building, but to validate protein-design algorithms as those implemented in the computer program *MUMBO* for their use in crystallography. The results we obtained so far are very encouraging. It appears obvious to us that the full potential this approach provides would become more evident if these algorithms were combined with other already existing algorithms for the tracing of the backbone or the automated picking of solvent molecule; for example, those implemented in *ARP*/*wARP* (Perrakis *et al.*, 1999). Combining these methods would generate the added benefit of using a detailed context-dependent description of the side-chain environment on one hand and at the same time through the concomitant adjustment of the protein backbone helping to overcome a severe limitation of side-chain packing algorithms, namely that of the fixed backbone trace.

An obvious parallel exists between model building during crystallographic refinement and *de novo* protein design, where side-chain packing and selection algorithms have proved very successful in the past. Here, we showed that upon introduction of an X-ray pseudo-energy the same algorithms can be applied successfully to the model-building step during the crystallographic refinement procedure.

We would like to thank the following people from University Erlangen-Nuremberg: Heinrich Sticht and Anselm Horn for many fruitful discussions regarding the layout of the force field, Andrea Thorn for help with the test cases and Benedikt Schmid for comments on the manuscript. We thank Kay Diederichs from the University of Konstanz for help with the parallellization of the computer code, for discussions and for critical reading of the manuscript.

References

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.
Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). *Science*, **253**, 164–170.
Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comput. Chem.* **4**, 187–217.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., 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.
Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L. Jr (2003). *Protein Sci.* **12**, 2001–2014.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dahiyat, B. I. & Mayo, S. L. (1997). *Science*, **278**, 82–87.
De Maeyer, M., Desmet, J. & Lasters, I. (2000). *Methods Mol. Biol.* **143**, 265–304.
Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). *Nature (London)*, **359**, 539–542.
Drenth, J. (1994). *Principles of Protein X-ray Crystallography*. Berlin: Springer-Verlag.
Dunbrack, R. L. Jr & Cohen, F. E. (1997). *Protein Sci.* **6**, 1661–1681.
Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). *Science*, **304**, 1967–1971.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Gohlke, H., Kuhn, L. A. & Case, D. A. (2004). *Proteins*, **56**, 322–337.
Goldstein, R. F. (1994). *Biophys. J.* **66**, 1335–1340.
Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). *J. Comput. Chem.* **24**, 232–243.
Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). *Nature (London)*, **358**, 86–89.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Ko, T.-P., Day, J., Greenwood, A. & McPherson, A. (1994). *Acta Cryst.* D**50**, 813–825.
Koga, Y., Caccia, N., Toyonaga, B., Spolski, R., Yanagi, Y., Yoshikai, Y. & Mak, T. W. (1986). *Eur. J. Immunol.* **16**, 1643–1646.
Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). *Nature Struct. Mol. Biol.* **11**, 371–379.
Kortemme, T., Morozov, A. V. & Baker, D. (2003). *J. Mol. Biol.* **326**, 1239–1259.
Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). *Science*, **302**, 1364–1368.
Lazaridis, T. & Karplus, M. (1999). *Proteins*, **35**, 133–152.
Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.
Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). *Nature (London)*, **423**, 185–190.
Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *Proteins*, **40**, 389–408.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.
Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.
Muller, Y. A., Ultsch, M. H. & de Vos, A. M. (1996). *J. Mol. Biol.* **256**, 144–159.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Neria, E., Fischer, S. & Karplus, M. (1996). *J. Chem. Phys.* **105**, 1902–1921.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000). *J. Comp. Chem.* **21**, 999–1009.
Pokala, N. & Handel, T. M. (2005). *J. Mol. Biol.* **347**, 203–227.
Ponder, J. W. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J., Wu, H., Choi, H. K. & Lynch, R. E. (1992). *J. Appl. Cryst.* **25**, 166–180.

Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). *Proteins*, **37**, *Suppl.* 3, 171–176.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Tronrud, D. E. (2004). *Acta Cryst.* D**60**, 2156–2168.
Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. G. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 3778–3783.
Weiss, M. S., Palm, G. J. & Hilgenfeld, R. (2000). *Acta Cryst.* D**56**, 952–958.