High-resolution structure of phenol hydroxylase and correction of sequence errors

Cristofer Enroth†

EMBL Hamburg Outstation, Notkestrasse 85, 22603 Hamburg, Germany

† Present address: Karo Bio AB, NOVUM, S-14157, Huddinge, Sweden. E-mail: cristofer.enroth@karobio.se.

Correspondence e-mail: cristofer.enroth@embl-hamburg.de

The crystal structure model of phenol hydroxylase has been corrected for 11 sequence errors and refined against new data to 1.7 Å resolution. The higher resolution data together with careful exploitation of non-crystallographic symmetry restraints and the use of many small groups for refinement of anisotropic displacement parameters resulted in a large decrease in the crystallographic R factor. The final crystallographic free R factor is 18.0%, which should be compared with the values of 27.8% for the previously published model (PDB code 1foh). The rebuilding and re-refinement procedure is described. A comparison with the previously published model was performed and possible biochemical implications are discussed. No large differences suggesting gross errors in the earlier model were found. The actual differences between these two models give an indication of the level of ambiguity and inaccuracy that may be found in a well refined protein model at 2.4 Å resolution.

1. Introduction

Phenol hydroxylase (PHOX) is an FAD-dependent oxidoreductase which catalyses the first oxidation step in the biodegradation of phenol by the soil-inhabiting yeast *Tricosporon cutaneum*. The biologically active unit of PHOX is a dimer of 2×664 amino acids. The cloning and sequence of PHOX were reported in 1992 (Kalin *et al.*, 1992). PHOX crystallizes in space group $P2_1$, with unit-cell parameters $a = 100.0$, $b = 150.9$, $c = 115.0$ Å, $\beta = 114.6°$ and two dimers in the crystallographic asymmetric unit. The solvent content is approximately 47% (Enroth *et al.*, 1994). The fourfold non-crystallographic symmetry breaks down as the two monomers in each dimer show two distinctly different conformations. Significant differences between the subunits include a large movement of residues 175–210 and 45–51, as well as a rotation of the isoalloxazine ring of the FAD cofactor (Enroth *et al.*, 1998). The two different conformations have been termed the 'in' and 'out' conformations, alluding to the position of the isoalloxazine ring with respect to the bound substrate in the active site.

During construction of site-directed mutants of phenol hydroxylase, 12 potential sequence errors were detected (Table 1; Xu *et al.*, 2001). One of these 12 potential errors had already been corrected, based on electron-density considerations, in the original crystal structure (Enroth *et al.*, 1998). The corresponding 11 positions in the phenol hydroxylase model (PDB code 1foh) were inspected in the original electron-density map. All 11 positions were corrected in all four subunits in the model. An initial refinement, using *REFMAC*5

Table 1
Data-collection parameters and data-reduction statistics.

Approximate crystal size (mm)	$1.0 \times 0.8 \times 0.6$		
Beamline/detector	BW7B/MAR345 IP		
Wavelength (Å)	0.8452		
No. images/$\delta\varphi$ (°)	395/0.5		
Wilson B factor (Å^2)	22.3		
Resolution range† (Å)	20.0–5.70	20–1.68	1.69–1.68
No. of unique reflections	8823	297900	3440
Completeness (%)	97.6	83.9	38.7
Redundancy‡	2.5	2.1	0.77
Linear R factor§¶	0.016	0.033	0.256
Square R factor§††	0.020	0.024	0.246
$I/\sigma(I)$	52	23	2.2

† Data were collected in a single rotation of the crystal. ‡ Redundancy was calculated as (total No. of observed reflections)/(total No. of possible unique reflections). § No σ cutoff was applied. Intensities were integrated with *DENZO*, merged and scaled with *SCALEPACK* (Otwinowski & Minor, 1997) and converted to structure factors with *TRUNCATE* from the *CCP*4 program package (Collaborative Computational Project, Number 4, 1994). ¶ $\sum(|I - \langle I \rangle|)/\sum(I)$. †† $\sum[(|I - \langle I \rangle|)^2]/\sum(I^2)$.

(Murshudov *et al.*, 1997, 1999; Winn *et al.*, 2001) against the original data, gave a significant drop in the conventional as well as in the free crystallographic R factor.

Traditionally, the refinement of anisotropic temperature factors has been confined to the realm of 'atomic resolution' models (better than 1.2 Å resolution), where the number of data points clearly exceeds the number of parameters of the model, including parameters for anisotropic temperature factors. However, as shown in this paper, if the data is of good quality and if non-crystallographic symmetry exists and is extensively utilized, careful refinement of anisotropic displacement parameters, as implemented in *REFMAC*5, will improve the model even at as low a resolution as 1.7 Å, as indicated by a significant drop in the free R factor. The TLS refinement thus offers the possibility of parameterizing anisotropic movement in the model without introducing too many new refinement parameters.

Furthermore, the final model contains new features, such as additional solvent molecules, main-chain shifts and altered side-chain conformations. A detailed comparison with the old model is discussed in this paper. Even small inaccuracies, as exemplified in the old 2.4 Å resolution model, may have a calamitous impact on, for example, molecular-dynamics simulations of the enzymatic function. Therefore, it is important to improve the molecular models as soon as new data is available, such as a corrected amino-acid sequence or better diffracting crystals.

2. Methods

2.1. Data collection and data quality

New crystals of the recombinant wild-type enzyme were prepared according to a previously published protocol (Enroth *et al.*, 1994, 1997). The availability of a larger image-plate detector (MAR 345) and an intense synchrotron X-ray beam made it possible to collect diffraction data to a nominal resolution of 1.68 Å compared with the previously published 2.4 Å. Data were collected at EMBL-Hamburg beamline BW7B (van Silfhout & Hermes, 1995) from one single flash-

Table 2
Real-space correlation coefficients for the non-corrected and corrected residues.

Residue	1foh	New	Correlation coeff.†	
			1foh	New
123	Histidine	Arginine	0.687	0.911
171	Aspartic acid	Glutamic acid	0.621	0.774
172	Histidine	Aspartic acid	0.296	0.692
186	Serine	Glycine	0.438	0.890
189	Histidine	Arginine	0.476	0.884
405	Histidine	Glutamine	0.818	0.924
406	Alanine	Proline	0.718	0.931
532	Serine	Alanine	0.624	0.890
544	Leucine	Arginine	0.652	0.804
549	Valine	Glycine	0.227	0.922
550	Serine	Alanine	0.644	0.924
265‡	Proline	Arginine	0.949	0.932

† Correlation coefficient for the side-chain atoms to the final $2F_o - F_c$ map, averaged over the four corresponding residues in the different subunits of the model. The correlation coefficients were calculated with the program *O*, using default parameters. ‡ Already corrected in 1foh in the PDB. It is included here for completeness. The correlation reported here for 1foh is that for an arginine residue.

frozen crystal. Data-collection and reduction statistics are shown in Table 1. The intensities were integrated with *DENZO* and scaled and merged with *SCALEPACK*, both from the *HKL* program package (Otwinowski & Minor, 1997). The data have a good signal-to-noise ratio: $I/\sigma(I)$ is 23 and 2.2 for overall data and the highest resolution shell, respectively. The overall R_{sym} is 3.3%. The completeness in the highest resolution shell is limited owing to a slight anisotropy of the diffraction. However, in the resolution bin 1.94–1.91 Å the completeness is above 90% and it rises to above 95% for all resolution bins below 2.10 Å resolution. The final refinement was performed with data to 1.7 Å resolution.

2.2. Model rebuilding

The refinement of a protein model the size of PHOX is a time-consuming task. The corrected amino acids were substituted into the model in proper rotamer conformations. One round of refinement resulted in an immediate and significant decrease in both the conventional and the free R factor. All corrected amino acids had a higher real-space correlation to the final map than their corresponding faulty counterparts in the old model. In the next step, new data were collected and proper rebuilding and refinement was carried out in several iterative steps. The corrected residues superimposed on the old faulty residues are shown together with the final electron-density map in Fig. 1 and the corresponding real-space correlation coefficients are presented in Table 2. An overview of the major steps in the rebuilding and refinement process is given in Table 3.

2.3. Anisotropic refinement

TLS refinement is a powerful tool to model anisotropic movement of domains and to account for anisotropy in the data (Winn *et al.*, 2001). For the purpose of TLS refinement, each subunit was divided into eight structural domains. The FAD cofactors were treated as a separate TLS groups. Each

Table 3
Some major events during the rebuilding and refinement of phenol hydroxylase.

Model	R (%)	R_{free} (%)	R.m.s.d. bond lengths (Å)	R.m.s.d. bond angles (°)	What happened to the model?
1foh	21.7	27.8	0.010	2.0	Previously published model
2	21.2	26.0	n.a.	n.a.	Corrected 4×11 sequence errors in original model, original 2.4 Å resolution data, no refinement
3	20.66	24.12	0.035	2.83	New data to 1.8 Å resolution included and NCS restraints completely released, loose stereochemical restraints
4	20.70	23.38	0.028	2.28	Minor rebuilding and addition of solvent waters (to a total of about 1900 waters), tighter stereochemical restraints
(5a)	26.26	29.59	n.a.	n.a.	All atomic B factors reset to 40 Å^2 before TLS refinement
5	17.90	21.36	0.016	1.64	TLS refinement in *REFMAC*5 (28 groups), waters ignored in TLS
(6a)	22.83	22.50	0.016	1.67	New data to 1.7 Å resolution (note: new R_{free} set), small change in unit-cell parameters
6	17.63	19.75	0.017	1.66	TLS refinement (40 groups), including waters in the TLS groups
7	15.70	17.99	0.012	1.39	New NCS restraints assigned, including waters (where applicable) using *WATNCS* and manual intervention

Table 4
A summary of the refinement statistics.

Resolution range (Å)	20–1.7	1.74–1.70
No. of unique reflections used	277056	11207
Reflections used for free R factor	14145	600
R factor (%)	15.70	23.5
Free R factor (%)	17.99	27.8
Model content		
No. of non-H atoms used in refinement	23726	
No. of flavin atoms	212	
No. of phenol atoms	28	
No. of chloride ions	4	
No. of water molecules	2688	
Ramachandran statistics† (%)		
Most favoured regions	90.7	
Additional allowed regions	8.9	
Generously allowed regions	0.2	
Disallowed regions	0.2	
Bad contacts per 100 residues	0.3	
R.m.s. deviation from ideality		
Bond lengths (Å)	0.012	
Bond angles (°)	1.40	
Average B factors‡ (Å^2)		
All protein atoms	11.2	
Main-chain atoms	10.0	
Side-chain atoms	12.5	
FAD	10.5	
Phenol	29.1	
Waters	19.7	
R.m.s. deviation in B factors for bonded atoms (Å^2)		
Main-chain atoms	1.09	
Side-chain atoms	2.82	

† As defined by *PROCHECK* (Laskowski *et al.*, 1993). ‡ Residual B factors after TLS refinement in *REFMAC*5 (Winn *et al.*, 2001).

TLS group also included a number of associated solvent water molecules. The TLS refinement is particularly powerful in models with non-crystallographic symmetry (NCS) because it parameterizes the anisotropic displacements of whole domains in the subunits, resulting in atomic isotropic displacement parameters (residual individual temperature factors) which are more suitable for inclusion in NCS restraints. The TLS groups of one subunit are shown in Fig. 2.

2.4. Treatment of non-crystallographic symmetry

Each of the four subunits of phenol hydroxylase in the crystallographic asymmetric unit experiences a somewhat different environment with respect to crystal-packing interactions. This gives rise to differences in the conformations of residues at the surface of the molecules, as well as differences in the positions of the surrounding water molecules, *i.e.* the fourfold NCS breaks down at these specific regions. Utilizations of a combination of fourfold, all possible threefold and all possible twofold NCS restraints are in principle useful to increase the number of restraints of the model. However, some of these possible groups of NCS restraints would contain very few atoms: a couple of solvent molecules, a residue or two or even only a part of a residue. Considering the potential bias introduced when assigning such restraint groups, it was decided for the current model to only use combinations of fourfold and twofold NCS restraints; furthermore, the twofold restraints were limited to those between subunits in the same overall conformation ('in' and 'out' conformation, respectively), thus reducing the number of restrained groups substantially while still including 2440 of the 2618 residues in the protein model in any of the NCS restraints. The total number of water molecules not involved in any NCS restraints is only 294 out of the total of 2688 modelled waters.

2.5. Avoiding model bias in map calculations

During the refinement, the crystallographic R and free R factors were monitored. In an attempt to avoid model bias, several different maps were calculated: a 'composite omit map' as defined in *CNS* (Brünger *et al.*, 1998), a map from an automatically built free-atoms model using *wARP* with 'mode molrep' (Perrakis *et al.*, 1999, 2001) and $F_o - F_c$, $2F_o - F_c$ as well as $3F_o - 2F_c$ maps using *REFMAC*5 and *FFT* from the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994). The total CPU time used to produce these maps using *CNS* and *wARP* was about 6 and 2 d, respectively, on a Compaq ES40 computer. As *wARP* would automatically build less than 1650 amino acids (out of the expected >2600), this model was discarded, while the map was kept and used for manual rebuilding with *O* (Jones *et al.*, 1991).

2.6. Treatment of ordered solvent

Water molecules were added to the model for $F_o - F_c$ electron-density peaks stronger than 3.5σ and if a reasonable hydrogen-bond geometry was present (within 2.5–3.4 Å distance from its bonding partner). All new solvent molecules were inspected manually with the *O* program before refine-

ment. More than 89% of the water molecules were included in the NCS restraints during refinement; the rest of the solvent molecules were found in the various interfaces between subunits. The geometry and chemical environment of each solvent position was also evaluated using the program *XPAND* (Kleywegt, unpublished program). One chloride ion was introduced per subunit after considering the chemical environment and the very low temperature factor assigned for the corresponding water molecule. The chloride ion is coordinated by three main-chain N atoms, residues 69–71 and two water molecules. Presumably, the ion originates from the

crystallization solution and has no biochemical significance, as it is situated more than 20 Å from the active site.

3. Results

3.1. Quality of the final model

The final model is excellent: it explains the data well, as shown by the very low free *R* factor for a model of this resolution, and it conforms to expected stereochemical parameters, as well as having few Ramachandran outliers. The
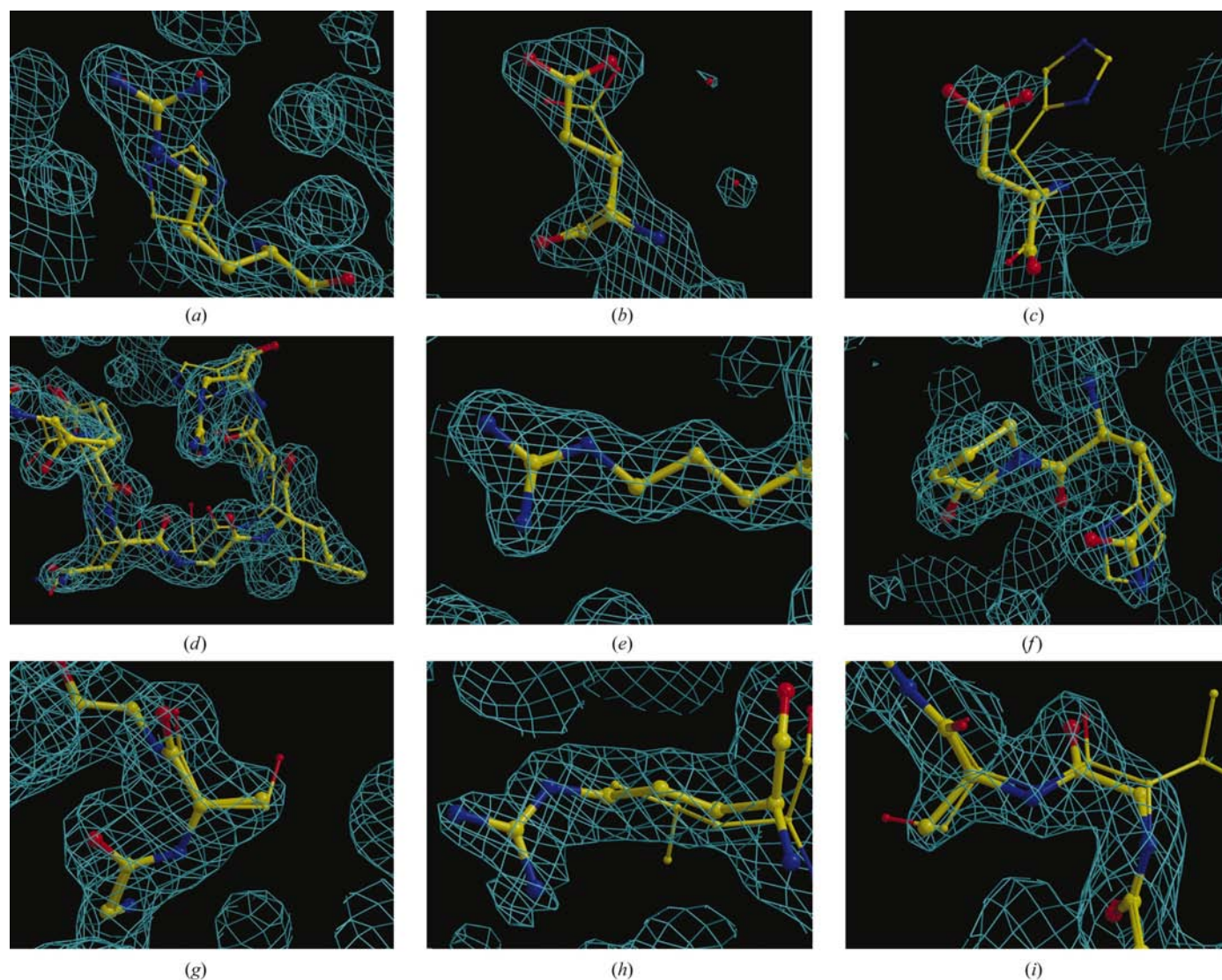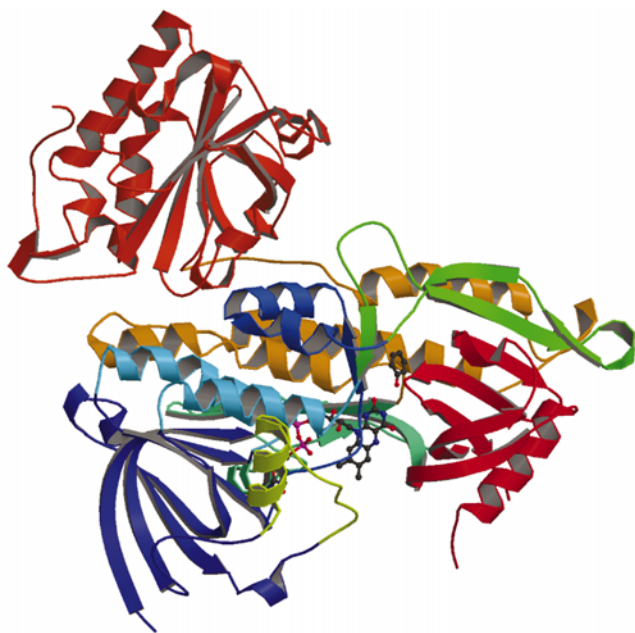

Figure 1
(*a*)–(*i*) show examples of electron densities at the changed side chains. The images were derived from subunit *A* (one of the 'open' subunits), except where stated otherwise. The new side chains are shown in a thicker representation overlaying the old side chain from 1foh. The density shown is $2F_o - F_c$ contoured at different levels between 1σ and 2σ for optimum clarity. This figure was prepared with *BOBSCRIPT* (Esnouf, 1999) and *Raster3D* (Merritt & Murphy, 1994). (*a*) His123 and two solvent water molecules were exchanged for Arg, with a real-space correlation coefficient (RSCC) of 0.920. (*b*) Asp171 was changed to Glu171, improving the RSCC from 0.781 to 0.855. (*c*) On the surface of the subunits, His172 was changed to Asp in weak density. (*d*) Ser186 (bottom, middle) and His189 (top, middle) were replaced by the much better fitting residues Gly186 and Arg189. (*e*) Pro265 was already corrected to Arg (on the basis of electron density) in the original model. Arg clearly explains the density better. (*f*) Density at His405 and Ala406 is better explained by Gln and Pro, respectively. (*g*) Ser532 mutated to Ala, an obvious improvement. (*h*) At position 544, an Arg fits better than a Leu. (*i*) Val549 and Ser550 mutated to Gly and Ala.

refinement and model statistics are presented in Table 4. This new model of phenol hydroxylase has 'inside' or 'better' than expected values for a 1.7 Å resolution model for all quality parameters evaluated by *PROCHECK* (Laskowski *et al.*, 1993).
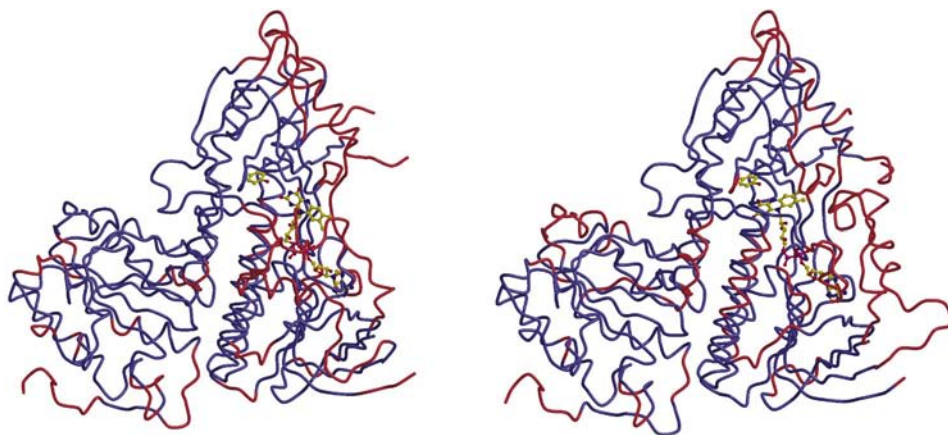
3.2. Comparison with the previously published model

A comparison of the new model with the old revealed some differences, besides the corrected sequence errors, in the

Figure 2
A figure showing the eight structural domains of a subunit which were treated as independent domains in the TLS refinement. FAD, which was treated as another separate TLS domain, is shown in ball-and-stick representation. This figure was prepared with *MOLSCRIPT* (Kraulis, 1991) and *Raster3D* (Merritt & Murphy, 1994).

conformations of small side chains. No large differences in the main-chain trace could be detected. All corrected side chains had a significantly lower real-space correlation coefficient in the final $2F_o - F_c$ map than the previous (erroneous) side chains. An overview of electron densities for the new side chains is shown in Fig. 1. A superposition of the old model on the new, performed with the program *ESCET* (Schneider, 2002), highlighting the significant differences in the Cα positions, is shown in Fig. 3. From the figure, it is immediately evident that the significant differences are mainly in loops and on the surface of the subunits. Furthermore, it can be seen that the differences are distributed over the whole model. There was not a specific correction ('error') in the old model that could be the cause of the large drop in free R factor, *i.e.* the old model was correct and complete (accurate) to the level of precision given by the lower resolution. The r.m.s.d. between the positions of the 2605 Cα atoms of the old and the new models is 0.275 Å, which is comparable to the expected error in atomic positions in a 2.4 Å resolution model. The observed positional differences are not just a reflection of the observed changes in unit-cell parameters (the new cell volume is 0.22% larger than in the former crystal, corresponding to unit-cell parameter differences of $\delta a = -0.03$, $\delta b = 0.27$, $\delta c = 0.04$ Å, $\delta\beta = -0.047°$), as a simple scaling of the coordinates did not improve the superposition of the two models.

The corrected residues that are the closest to the cofactor or to the active site are Gly186 and Arg189. These two residues are situated on the long flexible loop that moves in concert with the FAD cofactor during turnover. In the closed conformation subunits, these residues are about 6 Å from the methyl groups of the isoalloxazine ring and an Arg189 side-chain N atom makes a hydrogen bond to one of the phosphate O atoms of the FAD cofactor. After correction of the amino-acid sequence, several new ordered solvent water molecules could be placed in the vicinity of Arg189. In the open conformation, Gly186 and Arg189 point outwards towards the opposite subunit of the dimer. Residues Leu187 and Phe188 make dimer contacts. Gly186 and Arg189 have good electron density in both conformations.

4. Discussion

A protein model may be used for many purposes, *e.g.* for prediction of function, inhibitor design, structural comparisons, elucidation of catalytic mechanism, for use as a molecular-replacement search model *etc*, all of which require different precision and accuracy of the model. Before utilizing a structural model for whatever purpose, it is important to obtain a clear assessment of the quality and

Figure 3
Two superpositions of the two different subunits ('open' and 'closed' conformation, respectively) from one dimer of phenol hydroxylase. The red colour indicates areas where significant differences are found, while the blue colour indicates parts of the structure where only non-significant differences are found. The significance level was defined using *ESCET* (Schneider, 2002) for a model of this quality. Real differences are found in loop regions, on the surface of the subunits and in crystal contacts, *i.e.* in the less well defined parts of the low-resolution model. Owing to the slight differences in unit-cell parameters, the differences are somewhat exaggerated. The figure was prepared with *MOLSCRIPT* (Kraulis, 1991).

reliability of that model. A model refined against data to 1.7 Å resolution clearly has much better precision than a model refined against data to 2.4 Å. Of course, small errors in bond lengths represent large errors in bond energies and may completely disrupt any molecular-dynamics or quantum-mechanical simulation that the model could (or could not!) be used for. A minor error, such as a faulty side-chain rotamer or a missing/extra water molecule, will completely disrupt any attempt to predict function with a reasonable energy or spatial resolution. The coordinate-error estimate for this high-resolution model is 0.062 Å, based on maximum-likelihood weighting. Furthermore, if water molecules are implicated in the reaction mechanism and play a role in the catalysis, the placement of these waters is going to affect any attempt to simulate the reaction mechanism. Thus, it is important to investigate what kind of errors may be found in models built to different resolutions. In this study, we discuss the kind of inaccuracies that are found in a 2.4 Å model and which could be corrected by re-refinement to 1.7 Å resolution. At 1.7 Å resolution one may be confident that solvent molecules are placed in correct positions and have reasonable temperature factors.

Although no major structural differences are expected, it is important to re-evaluate crystallographic structures, such as that of phenol hydroxylase, when important technical advances have been made. Over the last 5 y, since the previous model was published, there have been several advances in the field of protein crystallography. Firstly, the commercial availability of larger image plates has made it feasible to collect data to higher resolution. Secondly, maximum-likelihood refinement methods and the possibility of refining anisotropic displacement parameters in a convenient fashion, as implemented in TLS refinement in *REFMAC*5 (Winn *et al.*, 2001), have significantly improved the way a model may be parameterized during refinement. Thirdly, new methods of comparing and validating protein models have been developed. The fact that phenol hydroxylase contains NCS, diffracts to high resolution and shows relatively large and well defined differences between the subunits makes it suitable as a test case, *e.g.* for the use of local density correlation maps (Kleywegt, 1999) or error-scaled difference distance analysis (Schneider, 2002) to validate differences between NCS-related subunits.

5. Conclusions

It is possible to refine a model at 1.7 Å resolution to a final free *R* factor below 18% using NCS restraints, TLS parameters and careful modelling. The new model has significantly higher precision, as indicated by a very small estimated coordinate error. It is also more accurate in explaining the data, as indicated by the low free *R* factor. The most powerful parameters

of the model, for improving the *R* factors, were the TLS tensors. Introducing the TLS parameters improved the free *R* factor from approximately 23 to 18%, while also improving the stereochemistry significantly.

As shown by this example it could be worthwhile to revisit and re-refine a crystallographic model even if no errors are expected in the original model. Any structural model may be significantly improved using the latest available refinement and rebuilding techniques. For the purpose of possible later improvements of a model it is important that crystallographers deposit not only the final models but also the structure factors in the public databases.

References

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.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Enroth, C., Huang, W., Waters, S., Neujahr, H., Lindqvist, Y. & Schneider, G. (1994). *J. Mol. Biol.* **238**, 128–130.

Enroth, C., Neujahr, H., Schneider, G. & Lindqvist, Y. (1998). *Structure*, **6**, 605–617.

Enroth, C., Waters, S., Neujahr, H., Lindqvist, Y. & Schneider, G. (1997). *Flavins and Flavoproteins XII*, edited by K. Stevenson, V. Massey & C. Williams, pp. 337–340. Alberta, Canada: University of Calgary Press.

Esnouf, R. M. (1999). *Acta Cryst.* D**55**, 938–940.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Kalin, M., Neujahr, H. Y., Weissmahr, R., Sejlitz, T., Johl, R., Fiechter, A. & Reiser, J. (1992). *J. Bacteriol.* **174**, 7112–7120.

Kleywegt, G. (1999). *Acta Cryst.* D**55**, 1878–1884.

Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Merritt, E. A. & Murphy, M. E. P. (1994). *Acta Cryst.* D**50**, 869–873.

Murshudov, G. N., Lebedev, A., Vagin, A. A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–255.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Perrakis, A., Harkiolaki, M., Wilson, K. S. & Lamzin, V. S. (2001). *Acta Cryst.* D**57**, 1445–1450.

Perrakis, A., Morris, R. M. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Schneider, T. R. (2002). *Acta Cryst.* D**58**, 195–208.

Van Silfhout, R. G. & Hermes, C. (1995). *Rev. Sci. Instrum.* **66**, 1818–1820.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst.* D**57**, 122–133.

Xu, D., Ballou, D. P. & Massey, V. (2001). *Biochemistry*, **40**, 12369–12378.