Journal of Computer-Aided Molecular Design, 8 (1994) 607–622
ESCOM

J-CAMD 258

An evaluation of molecular models of the cytochrome P450 *Streptomyces griseolus* enzymes P450SU1 and P450SU2

Julie A. Braatz*, Michael B. Bass** and Rick L. Ornstein***

Molecular Science Research Center, Pacific Northwest Laboratory, Battelle Memorial Institute, P.O. Box 999, Richland, WA 99352, U.S.A.

Received 15 December 1992
Accepted 12 April 1994

Key words: Homology modeling; Sequence alignment; Three-dimensional structure; Molecular mechanics

SUMMARY

P450SU1 and P450SU2 are herbicide-inducible bacterial cytochrome P450 enzymes from *Streptomyces griseolus*. They have two of the highest sequence identities to camphor hydroxylase (P450cam from *Pseudomonas putida*), the cytochrome P450 with the first known crystal structure. We have built several models of these two proteins to investigate the variability in the structures that can occur from using different modeling protocols. We looked at variability due to alignment methods, backbone loop conformations and refinement methods. We have constructed two models for each protein using two alignment algorithms, and then an additional model using an identical alignment but different loop conformations for both buried and surface loops. The alignments used to build the models were created using the Needleman–Wunsch method, adapted for multiple sequences, and a manual method that utilized both a dot-matrix search matrix and the Needleman–Wunsch method. After constructing the initial models, several energy minimization methods were used to explore the variability in the final models caused by the choice of minimization techniques. Features of cytochrome P450cam and the cytochrome P450 superfamily, such as the ferredoxin binding site, the heme binding site and the substrate binding site were used to evaluate the validity of the models. Although the final structures were very similar between the models with different alignments, active-site residues were found to be dependent on the conformations of buried loops and early stages of energy minimization. We show which regions of the active site are the most dependent on the particular methods used, and which parts of the structures seem to be independent of the methods.

INTRODUCTION

The cytochromes P450 are a large family of enzymes, capable of oxidizing a variety of substrates from alkanes to polycyclic aromatic hydrocarbons [1,2]. Of this large family of enzymes,

*Present address: Department of Chemistry and Biochemistry, University of California at Santa Cruz, Santa Cruz, CA 95064, U.S.A.
**Present address: Amgen Inc., 1840 Dehavilland Drive, Thousand Oaks, CA 91320-1789, U.S.A.
***To whom correspondence should be addressed.

detailed three-dimensional (3D) information is known for only one enzyme, camphor hydroxylase (P450cam; P450 101) from *Pseudomonas putida* [3]. The atomic structure of this enzyme has provided insight into the reaction and substrate specificity of this particular enzyme. Studies have yielded detailed information about the atomic interactions between camphor hydroxylase and the substrate [3], as well as several other substrate analogues [4]. (While the present manuscript was being revised in response to referee comments, the 3D structure of the hemoprotein domain of cytochrome P450BM-3 was published [Ravichandran et al., Science, 261 (1993) 731–736].)

Recent attempts to alter the substrate specificity [5,6] or the high-spin/low-spin equilibrium [7] of specific enzymes in the P450 superfamily have made use of the similarity between the gene sequences. To aid in the rational redesign of substrate specificity of the eukaryotic P450 enzymes, 3D models have been constructed of the active site of aromatase [8], the entire proteins of sterol 14-α-demethylase [9], sterol 17-α-hydroxylase [10], and 3-methylcholanthrene hydroxylase (P450IA1) [11]. Each of these models used a single-method, multiple-sequence alignment to generate the correspondence between residues in the model and residues in P450cam. To date, homology models of members of the cytochrome P450 family have been limited to one choice of sequence alignment and one choice of model refinement. In this study, we address the variability of the final structure due to sequence alignment, loop building, and methods for structural refinement. How the sequences are aligned can dramatically affect the correspondence between residues in the modeled sequence and the known structure. This is especially true for regions of low sequence identity, where the alignment algorithms perform poorly. In the cytochrome P450 superfamily this is rather troublesome, as the substrate binding pocket displays the lowest identity of all the regions in the coding sequence [12]. Consequently, the regions of the protein most important to understanding substrate specificity are those that possess the least confidence in the sequence alignment. The way in which the loops are modeled can also result in a large difference in the 3D model. In the sequences studied here, there are insertions in regions bordering the active site. Different assumptions as to the backbone conformation in these regions may dramatically affect the predicted enzyme–substrate interactions. Furthermore, the method used to refine the preliminary model can also affect the final structure. In the approach we have used, no special care is taken to assure adequate packing around the side chains. The methods employed to relieve the poor 3D contacts in the preliminary structure will also have an effect on the local backbone structure. These effects will be greatest in the regions of lowest identity (e.g., the active site).

We have modeled the 3D structures of cytochrome P450SU1 (P450 105A1; hereafter called SU1) and P450SU2 (P450 105B1; hereafter called SU2) from *Streptomyces griseolus* [13]. Each of these enzymes is induced by sulfonylurea herbicides; they are known to catalyze the hydroxylation and dealkylation of chlorsulfuron [14]. The sequence homologies of these two sequences are much closer to the sequence of P450cam than any of the sequences for the eukaryotic cytochromes (they have 25% sequence identity to P450cam and 42% sequence identity with each other). However, SU1 and SU2 are herbicide-induced enzymes that hydroxylate a class of substrates [14] quite different from the small bicyclic compounds that P450cam acts on; they will likely have quite different substrate binding pockets. In previous studies in which molecular models were constructed for members of the P450 superfamily, no variants in the proposed structure were addressed. It is therefore difficult to evaluate how these models were dependent on the modeling process. We compare models built by several different methods of sequence alignment, loop building and structural refinement. We describe the dependence of the various

structural models on the alignment and modeling methods. This provides some measure of the level of confidence in the structures presented here.

METHODS

Sequence alignment

The P450cam [15], SU1 and SU2 sequences [13] were aligned using four alignment programs: the automatic methods GAP, BESTFIT and PILEUP (all found in the Wisconsin Genetics Computer Group package [16]), and a visual method, ISSC [17]. The alignment program GAP uses the Needleman–Wunsch algorithm, which finds the best global alignment between two sequences. The BESTFIT program uses the Smith–Waterman algorithm, which aligns sequences with emphasis on local regions of high similarity. PILEUP is a multiple-sequence alignment program that uses a modified Needleman–Wunsch algorithm. Here the sequences are iteratively aligned pairwise, starting with the most similar. All of these methods use the PAM 250 weighting matrix [18]. The ISSC algorithm incorporates a windowing algorithm, which is used to create a search matrix. The values of the search matrix are a combination of the Dayhoff matrix [18] and five residue-based physicochemical characteristics [19].

One of the tenants of homology modeling is that insertions or deletions (hereafter called 'gaps') in a sequence should not occur within secondary structures. However, the methods of sequence

TABLE 1
COMPARISON OF SEQUENCE ALIGNMENT METHODS

Method	Score[a]		Gaps		Significance[b]	
	SU2	SU1	SU2	SU1	SU2	SU1
Original[c]						
GAP	174.9	192.2	9	5	15.02	27.74
BESTFIT	175.3	192.2	9	5	15.25	27.74
ISSC	178.8	166.5	8	8	15.81	19.91
PILEUP	187.9	205.5	7	5	17.65	31.80
Edited[d]						
GAP	166.6	189.9	9	5	13.33	26.42
BESTFIT	167.1	189.9	9	5	13.55	26.42
ISSC	178.5	160.1	7	7	15.75	17.95
PILEUP	186.6	206.0	5	5	17.34	31.95

[a] Scores are the raw score for each alignment using the Dayhoff matrix with the default gap weight and default length weight penalties.

[b] In order to ascertain the degree of relatedness of P450cam with P450SU1 and P450SU2, the following procedure was used, as outlined by Doolittle [31]. The alignment score for 20 different computer-generated shuffled (i.e., scrambled or randomized sequences with the same composition and length) SU1 and SU2 sequences with P450cam were determined and averaged and a standard deviation computed. The alignment score obtained with the genuine sequence is compared with the mean shuffled score, and the number of standard deviations above (or below) is noted. Scores that are three or more standard deviations above the shuffled mean score can reasonably be expected to be related; we report here the number of standard deviations above the mean score ('significance').

[c] Original sequence alignment obtained from each method, without editing as described in footnote d.

[d] Adjustment in the sequence alignment to remove insertions and deletions from α-helices (see text in the Sequence alignment section of the Methods).

alignment used here do not consider secondary structures. For instance, in the SU2 sequence, gaps were placed in two helices by PILEUP and in four by GAP, BESTFIT and ISSC. Gaps in β-sheets 3 and 5 were placed by all of the alignment methods, except PILEUP. To maintain the α-helices, gaps were moved to the outside of the helices without greatly affecting the alignment scores. In doing this, the SU2 GAP and BESTFIT alignments became identical. Table 1 lists the statistical scores for the four alignments of the two proteins with P450cam, including the edited alignments.

Models were built using the edited PILEUP and ISSC alignments for both SU1 and SU2 (Fig. 1A,B). A third model of SU1 was built using the PILEUP alignment to investigate the differences in structure arising solely from alternate conformations of both surface and buried loops. The ISSC and PILEUP alignments were chosen for 3D modeling, because the GAP and BESTFIT alignments only differed from the ISSC alignments by gap placements in the loop regions. These differences would have been lost when the models were built, resulting in identical models. One

```
                                                                  A
Cam     1 --TTETIQSN ANLAPLPPHV PEHLVFDFDM YNPSNLSAGV QEAWAVLQES
PILE    1 MTDTATTPQT TDAPAFPSNR SCPYQL-PDG YAQLRDTPGP LHRVTLYDGR
ISSC    1 ---------- MTDTATTPQT TDAPAFPSNR SCPY----QL PDGYAQLRDT

                                        B                    B'
Cam    49 NVPDLVWTRC NGGHWIATRG QLIREAYEDY RHFSSECPFI PREAGEAYD-
PILE   50 QA--WVVTKH EAARKLLGDP RLSSNRTDDN FPATSPRFEA VRESPQAF--
ISSC   37 PGP-LHRVTL YDG-----RQ AWVVTKHEAA RKLLGDPRLS SNRTDDNFPA

                                                C
Cam    98 ----FIPT-- ------SMDP PEQRQFRALA NQVVGMPVVD K-LENRIQEL
PILE   90 -------I-- ------GLDP PEHGTRRRMT ISEFTVKRIK G-MRPEVEEV
ISSC   81 TSPRFEAVRE SPQAFIGLDP PEHGTRRRMT ISEFTVKRIK GMRPEVEEVV

          D                             E                      F
Cam   135 ACSLIESLRP QGQC-NFTED YAEPFPIRIF MLLAGLPEED IPHLKYLTDQ
PILE  130 VHGFLDEMLA AGPTADLVSQ FALPVPSMVI CRLLGVPYAD HEFFQDASKR
ISSC  131 HGFLDEMLAA GPTA-DLVSQ FALPVPSMVI CRLLGVPYAD HEFFQDASKR

                                    G                  H
Cam   184 M-TRPDGSMT FAEAKEALYD YLIPIIEQRR QKPGTDAISI VANGQVNGRP
PILE  180 LV-QSTDAQS ALTARNDLAG YLDGLITQFQ TEPGAGLVGA LVADQLANGE
ISSC  180 L-VQSTDAQS ALTARNDLAG YLDGLITQFQ TEPGAGLVGA LVADQLANGE

                          I                          J
Cam   233 ITSDEAKRMC GLLLVGGLDT VVNFLSFSME FLAKSPEHRQ ELIQRPERIP
PILE  229 IDREELISTA MLLLIAGHET TASMTSLSVI TLLDHPEQYA ALRADRSLVP
ISSC  229 IDREELISTA MLLLIAGHET TASMTSLSVI TLLDHPEQYA ALRADRSLVP

          K
Cam   283 AACEELLRRF SL--V--ADG RILTSDYEFH GVQLKKGDQI LLPQMLSGLD
PILE  279 GAVEELLRYL AI--ADIAGG RVATADIEVE GHLIRAGEGV IVVNSIANRD
ISSC  279 GAVEELLRYL AIADI--AGG RVATADIEVE GHLIRAGEGV IVVNSIANRD

                                                        L
Cam   329 ERENACPMHV DFSRQKVSHT TFGHGSHLCL GQHLARREII VTLKEWLTRI
PILE  327 GTVYEDPDAL DIHRSARHHL AFGFGVHQCL GQNLARLELE VILNALMDRV
ISSC  327 GTVYEDPDAL DIHRSARHHL AFGFGVHQCL GQNLARLELE VILNALMDRV

Cam   379 PDFSIAPGAQ --IQHKSGIV SGVQALPLVW
PILE  377 PTLRLAVPVE QLVLRPGTTI QGVNELPVTW
ISSC  377 PTLRLAVPVE --QLVLRPGT TIQGVNELPVTW
```

Fig. 1A. Alignment of P450cam with SU1 using the PILEUP and ISSC alignment programs. The α-helices of P450cam and the corresponding residues in SU1 are outlined.

of the largest alignment differences between the ISSC and PILEUP SU1 alignments is the B'-helix. This helix is of special interest, because it forms one of the borders of the active site. The B'-helix of SU1 comprises very different residues in the ISSC and PILEUP alignments: 71–78 in ISSC and 88–95 in PILEUP.

One of the unique features of cytochrome P450cam are the three *cis*-prolines: residues 89, 100 and 106 [3]. These are found in the B'-helix, the subsequent loop and the C-helix, respectively. In a study of 35 P450 sequences [19], Pro^{89} and Pro^{100} are not conserved, but Pro^{106} is in over half of the sequences. In seven of the eight alignments of SU1 and SU2, Pro^{106} is conserved (this is residue 101 in SU1 and SU2), and in the other alignment it is misaligned by one residue (the ISSC alignment of SU2). In the latter alignment, Pro^{93} is aligned with P450cam 100; however, this is a region of low homology with many insertions and deletions in the alignment with SU1 and SU2. We are unable to adequately determine the conformation around this proline, since both conformations can be adequately modeled into this region. Therefore, we have chosen to model

```
                                                                    A
Cam    1 --TTETIQSN ANLAPLPPHV PEHLVFDFDM YNPSNLSAGV QEAWAVLQES
PILE   1 -MTTAERTAP PDALTVPASR APGCPFDPAP DVTEAARTEP VTRATLWDGS
ISSC   1 ---------- ---MTTAERT APPDALTVPA SRAPGCPFDP APDVTEAART
                                    B                       B'
Cam   49 NVPDLVWTRC NGGHWIATRG QLIREAYEDY R----HFSSE CPFIPREAGE
PILE  50 SC--WLVTRH QDVRAVLGDP RFSADAHRTG F----PFLTA GGREIIGTNP
ISSC  38 EPVTRATLWD GSSCWLVTRH QDVRAVLGDP RFSADAHRTG FPFLTAGGRE
                                    C                       D
Cam   95 AYDFIPTSM- --DPPEQRQF RALANQVVGM PVVDK----- -LENRIQELA
PILE  94 TF----LRM- --DDPEHARL RRMLTADFIV KKVEA----- -MRPEVQRLA
ISSC  88 IIGTNPTFLR MDDPEHARLR RMLTADFIVK KVEAMRPE-- -VQRLADDLV
                                    E                       F
Cam  136 CSLIESLRPQ GQC-NFTEDY AEPFPIRIFM LLAGLPEEDI PHLKYLTDQM
PILE 131 DDLVDRMTTG RTSADLVTEF ALPLPSLVIC LLLGVPYEDH AFFQERSRVL
ISSC 135 DRMTTGRTSA ----DLVTEF ALPLPSLVIC LLLGVPYEDH AFFQERSRVL
                                    G                   H
Cam  185 T--RPDGSMT FAEAKEALYD YLIPIIEQRR QKPGTDAISI VANGQVNGRP
PILE 181 LT-LRSTPEE VRAAQDELLE YLARLARTKR ERPDDAIISR LVA----RGE
ISSC 181 LT-LRSTPEE VRAAQDELLE YLARLARTKR ERPDDAIISR LVA----RGE
                                I                   J
Cam  233 ITSDEAKRMC GLLLVGGLDT VVNFLSFSME FLAKSPEHRQ ELIQRPERIP
PILE 226 LDDTQIATMG RLLLVAGHET TANMTALSTL VLLRNPDQLA RLRAEPALVK
ISSC 226 LDDTQIATMG RLLLVAGHET TANMTALSTL VLLRNPDQLA RLRAEPALVK
          K
Cam  283 AACEELLRRF SLV--ADG-- RILTSDYEFH GVQLKKGDQI LLPQMLSGLD
PILE 276 GAVEELLRYL TIVHNGVP-- RIATEDVLIG GRTIAAGEGV LCMISSANRD
ISSC 276 GAVEELLRYL TIV--HNGVP RIATEDVLIG GRTIAAGEGV LCMISSANRD
                                                    L
Cam  329 ERENACPMHV DFSRQKVSHT TFGHGSHLCL GQHLARREII VTLKEWLTRI
PILE 324 AEVFPGGDDL DVARDARRHV AFGFGVHQCL GQPLARVELQ IAIETLLRRL
ISSC 324 AEVFPGGDDL DVARDARRHV AFGFGVHQCL GQPLARVELQ IAIETLLRRL

Cam  379 PDFSIAPGAQ --IQHKSGIV SGVQALPLVW
PILE 374 PDLRLAVPHE EIPFRGDMAI YGVHSLPIAW
ISSC 374 PDLRLAVPHE --EIPFRGDM AIYGVHSLPI AW
```

Fig. 1B. Alignment of P450cam with SU2 using the PILEUP and ISSC alignment programs. The α-helices of P450cam and the corresponding residues in SU2 are outlined.

this peptide bond in the trans conformation, since in model systems this conformation is generally more stable.

During the review of the manuscript, the following sequence errors were noticed by one of the referees. In the model of SU1, Gln^{9} was incorrectly modeled as a glutamate. In the model of SU2, three residues were incorrectly modeled. Glu^{38} and Glu^{125} were modeled as glycines, and Glu^{384} was modeled as a glutamine. Each of these residues is on the surface of the protein and at least 21 Å from the center of the active site or the center of the putative ferredoxin binding site. Hence, we do not expect these errors to have a significant impact on the results that follow.

Structural refinement

Model construction

The 3D models were constructed using the InsightII/Homology software (v. 2.0) by Biosym Technologies. Each sequence was divided up into α-helices and loops; the loops comprising all turns, β-sheets and random structure as inferred from alignment with the P450cam sequence. The helices and loops free of gaps were constructed three-dimensionally by using the backbone coordinates of the corresponding P450cam structure and extending the correct side chain. The backbone conformation for the loops containing gaps was obtained by searching the Brookhaven Protein Data Bank (PDB) [20] and selecting a backbone conformation that fit well in the structure. At this junction, no consideration was made for side-chain–side-chain interactions; large initial steric repulsions of the side chains created in this way were removed later by energy minimization. To complete the models, the P450cam crystallographic waters were added. This was done because the location of key buried waters might be conserved between closely related cytochromes P450. Only one water molecule overlapped with any protein atoms; it was removed prior to energy minimization. Hydrogen atoms were introduced into the models using only bond length and angle criteria. The dihedral angles of polar hydrogens were later rotated to form the greatest number of hydrogen bonds using the NETWORK program [21]. The first nine residues in P450cam lack coordinates, therefore all of the corresponding SU1 and SU2 residues (and any preceding residues) aligned to them were built as an extended chain.

Model minimization

Once built, the models were energy minimized using Discover (v. 2.7 and 2.8, Biosym), utilizing steepest descent and conjugate gradient methods, a distance-dependent (r) dielectric, and explicit treatment of all atoms. Parameters for the Lennard-Jones terms and partial atomic charges were obtained from the consistent valence force field [22,23]. Heme parameters were taken from our earlier work [24]. Model calculations used a net neutral approximation for physiologically ionic groups, unless specified otherwise. This type of model system and parameter set has been shown to be rather successful compared to even much more expensive explicit water calculations [25].

All models were minimized without the substrate docked into the active site. The minimizations were continued only until unrealistically long bonds created in the modeling process were shortened. This was a maximum gradient of 1 kcal mol^{-1} $Å^{-1}$ for all of the models. The total energy of the systems (protein and waters) ranged from −1265 to 1607 kcal after energy minimization. Evaluating the quality of these structures based on the energy values is not warranted. Since the structures were minimized to a maximum gradient of 1.0 kcal $Å^{-1}$, a significant change in

energy may still occur if further minimization were applied. In addition, the energies do not necessarily reflect how well modeled the active sites are. Loops far from the core of the protein may make large contributions to the energy, and therefore account for differences in total energy. Measuring the energies of only the active-site regions was not possible, because each contained a different number of atoms. Obviously, the above minimization criterion will not result in a fully minimized structure. In fact, we do not want to too strongly encourage the active site to repack, since it is devoid of any substrate or solvent during the minimization. Hopefully this criterion is a reasonable balance between achieving an approximate, globally packed structure and preserving the active-site void space. The models were energy minimized using the following four schemes.

The first method was a basic minimization that allowed the model to minimize in the following stages: first the hydrogens only, then the side chains followed by the backbone atoms of the loops, then the entire protein. At each point in the minimizations, all atoms that were not free to move were tethered using a 2000 kcal mol^{-1} $Å^{-1}$ force constant.

The second minimization scheme was performed to explore the question of whether allowing the least conserved α-helices to energy minimize with the loops, before the entire protein, would lead to a markedly different structure than that resulting from the first minimization. To do this, the SU1 and SU2 helices were divided in half, at similarity values of 43 and 50%, respectively. The rms deviations between these minimized structures and the final structures from the first minimizations were very low (0.488 and 0.309 Å), the differences being only in surface turns, a region not expected to be well modeled.

The third minimization scheme repeated the first, but substituted physiological charges on the glutamate, aspartate, lysine and arginine residues. This resulted in a net negative charge on both proteins, −18 on SU1 and −6 on SU2. This minimization scheme resulted in larger differences from the first minimization scheme than the second (rms deviations of 1.22–2.24 Å), but still insufficient to give a significantly different structure.

The fourth minimization was performed to attempt to change the packing of the models to more closely approximate the packing in natural proteins. QPACK [26] was used to find the volumes of the residues in the unminimized models. The program compares the volume of each residue in the structure to the ideal volumes of the amino acids and computes percentages of the ideal volumes. The unminimized model with the lowest volumes had residues with fractions of ideal volumes as low as 9% (this occurred because the side-chain atoms were added without taking the surrounding atoms into account). We tried to correct these deficiencies by allowing residues with 90% or less of the ideal volume to minimize throughout the minimization scheme. This was done by using the first minimization method, modifying it to allow the residues with small volumes to minimize after the hydrogens, and to continue to minimize throughout. The minimizations resulted in all of the residues in the models having more than 75% of the ideal volume. For reference, the smallest percentage of ideal volume in the P450cam crystal structure is 77%, and the smallest in the minimized P450cam structure is 75%.

The first three minimization schemes did not result in significant perturbations of side-chain or backbone positions in the active site. The QPACK-aided minimizations, on the other hand, resulted in changes in the active-site regions of the models. In the PILEUP model of SU2, Arg^{97}, previously hydrogen bonded to one of the propionate groups of the heme, shifted 180°. After minimization, the side chain was extended toward the surface of the protein. In the ISSC model of SU2 the changes are smaller, with Arg^{184} forming a hydrogen bond to Asp^{390} and Thr^{111} moving

towards the center of the active site, reducing the volume of the active site. In the ISSC model of SU1, Phe^{78} rotated about 45°, with a concurrent shift of the backbone atoms in this region.

Model differences due to loop construction

An additional model of SU1, using the PILEUP alignment, was built to explore the variability that can occur solely from the loop backbone conformations. Four loops in the protein, 25–29, 47–54, 289–295 and 379–389, were built using different backbone conformations from the ones chosen in the first PILEUP model. This was done by shifting the position of the loop by one or two residues to obtain new choices for the backbone conformations from the structural database. Three of the four loop regions were on the surface, and no significant differences occurred in the regions due to the loop conformations. The other loop, 290–297, is in the interior and includes active-site residues. This loop has a two-residue insert. In this case, the choice of a different conformation resulted in a significant difference in a large region of the protein. The surrounding area was perturbed, including a 6 Å shift of Phe^{95}. The alternate conformation of loop residues 290–297 pushed the side chain of Phe^{95} into the active site.

RESULTS AND DISCUSSION

To evaluate each model, structural functionality was addressed. This included the critical ferredoxin, heme, and substrate binding regions. The ferredoxin binding site is a highly conserved feature, necessary for electron transfer; the heme binding site must contain residues to neutralize the charge from the heme propionates; and the active site should show complementarity to the substrate, a sulfonylurea herbicide. Each of these features was examined as tests of the models.

Ferredoxin binding site

The P450cam, SU1 and SU2 systems all consist of an induced cytochrome P450, a ferredoxin, and ferredoxin reductase [3,14]. In P450cam, Arg^{72} and Lys^{344} have been proposed to be involved in ferredoxin binding [27]. In the SU1 and SU2 models, only the ISSC alignment of SU2 aligns a positively charged residue with the P450cam residues; the rest are aliphatic residues. As shown in Fig. 1A and B, Arg^{72} of P450cam is aligned with Val^{54} and Ser^{71} in SU1 and with Arg^{68} and Ala^{71} in SU2. Ala^{342} in SU1 and Ala^{339} in SU2 correspond to P450cam Lys^{344} in the alignments. However, in all of the models there are positively charged residues in the region of these aliphatic residues. In the SU1 models, Arg^{340} and Arg^{343} are near Ala^{342}, and Lys^{56} and Arg^{73} are near Val^{54} and Ser^{71}. In the SU2 models, Arg^{337}, Arg^{340} and Arg^{341} are near Ala^{342}, and Arg^{68} is near Ala^{71}. Thus, the presence of positive charge is still conserved. Since each inducible cytochrome P450 may have a slightly different ferredoxin [27], it is not surprising that the exact residues are not conserved, but that the general features of a positively charged binding site are.

Heme binding site

The heme propionates represent charges that are buried within the interior of the protein and stabilize the charge of the heme group. If this feature of the cytochromes P450 is conserved, then specific residues should also be present in the SU1 and SU2 structures. P450cam has five residues that hydrogen bond to the two propionate groups: Arg^{299}, Asp^{297}, Gln^{322}, Arg^{112} and His^{355}. Asp^{297} is unusual; it is thought to be protonated in P450cam [3], thus giving it a neutral charge. Of the

TABLE 2
RESIDUES THAT INTERACT WITH THE PROPIONATES IN P450cam AND THEIR EQUIVALENT RESIDUES IN THE MODELS

P450cam	SU2		SU1	
	ISSC	PILEUP	ISSC	PILEUP
Arg^{112}	Arg^{108}	Arg^{107}	Arg^{108}	Arg^{108}
Asp^{297}	Asn^{290}	Val^{292}	Gly^{295}	Gly^{295}
Arg^{299}	Arg^{294}	Arg^{294}	Arg^{297}	Arg^{297}
Gln^{322}	Ile^{317}	Ile^{317}	Asn^{320}	Asn^{320}
His^{355}	His^{350}	His^{350}	His^{353}	His^{353}

residues interacting with the heme propionates in P450cam, Arg^{299}, Arg^{112} and His^{355} are conserved in the SU1 and SU2 sequences, while Asp^{297} and Gln^{322} are not. Within the superfamily, Arg^{112} is conserved, Gln^{322} is found in about one-third of the sequences, and Arg^{299} is mostly replaced by a proline, with arginine appearing only occasionally [19]. His^{355} is replaced by an arginine in all 35 sequences in the study, and Asp^{297} is not conserved at all. This is not surprising, considering the unusual pKa suggested for this residue [3]. Examining the alignments of the SU1 and SU2 sequences, the amino acids corresponding to Asp^{297} are asparagine, valine and glycine. Gln^{322} is replaced by an isoleucine in the SU2 sequence and by an aspartate in the SU1 sequence. Table 2 summarizes the five residues that interact with the heme propionates in P450cam and the amino acids in the alignments that correspond to them. Figure 2 shows the heme binding site of the SU2 PILEUP model.

In addition to the conserved residues, there is one additional propionate-interacting group in most of the models. His^{103} is aligned with P450cam Glu^{108} in all of the alignments, except the ISSC alignment of SU2. In P450cam Glu^{108} is in the region of one of the heme propionates, but the side chain is pointing in another direction. However, in all of the models, except the ISSC model of SU2, His^{103} interacts with one of the propionate oxygens.

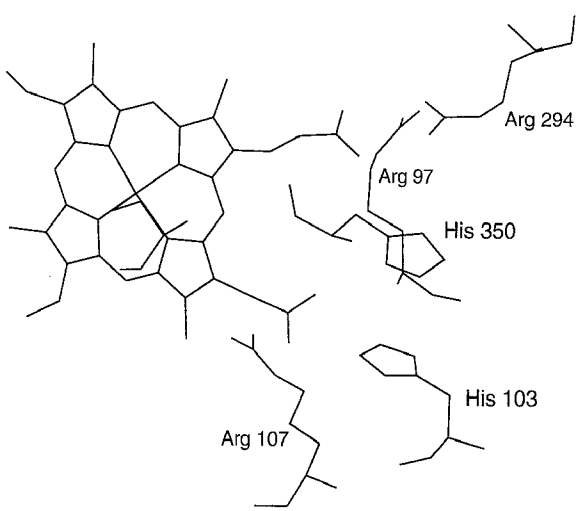

Fig. 2. Heme and residues interacting with the heme propionate groups of the PILEUP model of SU2.

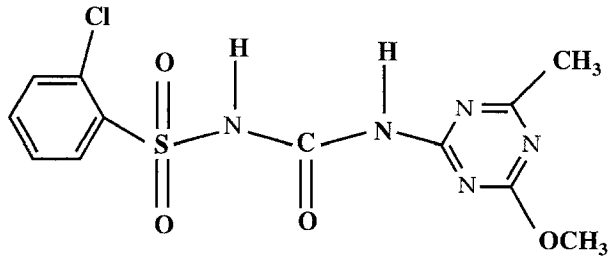

Fig. 3. Chlorsulfuron, one of the substrates of the inducible cytochromes SU1 and SU2.

Substrate binding site

Apart from these important regions in the models, we examined the substrate complementarity of the active site. To evaluate the active site we used chlorsulfuron (Fig. 3), a sulfonylurea herbicide which is one of the major substrates of both SU1 and SU2 [14]. The active-site residues are better conserved in the SU1 models than in the SU2 models; this is not surprising, considering the higher number of identities between P450cam and SU1 (28%). Chlorsulfuron was docked into the active sites of each of the refined models and the steric interactions and hydrogen bonding between the substrate and enzyme were analyzed. The substrate was placed in the active sites with the reactive methoxy group about 5 Å from the heme*, allowing room for molecular oxygen to bind, and avoiding the region where the oxygen is proposed to interact with a threonine. In SU2 this threonine is Thr^{245}, in SU1 it is Thr^{248}, and in P450cam it is Thr^{252}. The models were not energy minimized with the substrate in the active site. Thus, the active sites are not optimized for hydrogen bonding between the substrate and polar active-site residues. This was done so that the models would not be biased towards accommodating the substrate, and it could be determined whether or not the modeling methods would give rise to a reasonable structure that would bind substrate.

The active sites of the two SU2 models contain the same residues, with one notable exception: Arg^{86} and Phe^{95} are in the active site of the PILEUP model, but not in the ISSC model. The residues in the ISSC model in these positions are Asn^{92} and Arg^{97}, respectively. This is caused by the difference in the alignments of the B'-helix (Fig. 1B). Although the backbone structures of the models are similar in this region, different residues are in the active site and surrounding areas. The presence of arginines might provide an important contribution to substrate binding in the upper region of the active site, neutralizing the charge of the sulfone group of chlorsulfuron. After positioning the substrate in the active site, Arg^{86} in the PILEUP model hydrogen bonds with the sulfone group of chlorsulfuron (Fig. 4), providing some stabilization of the negative charge. Arg^{97} in the ISSC model is not in a position to interact with the sulfone group; its position is closer to the urea moiety of the substrate. The side chain points away from the substrate and has nonstandard bond angles. This is an artifact of the modeling process. We do not know the best position of this side chain; it could interact with the substrate. Arg^{236} is in the active site of both models and able to interact with the sulfone group if rotated in towards the center of the active site and the substrate (Fig. 4, rotated in towards the substrate). This was done to show probable positive interactions between the protein and substrate. There are no

*In this binding mode, the expected reaction would be dealkylation.

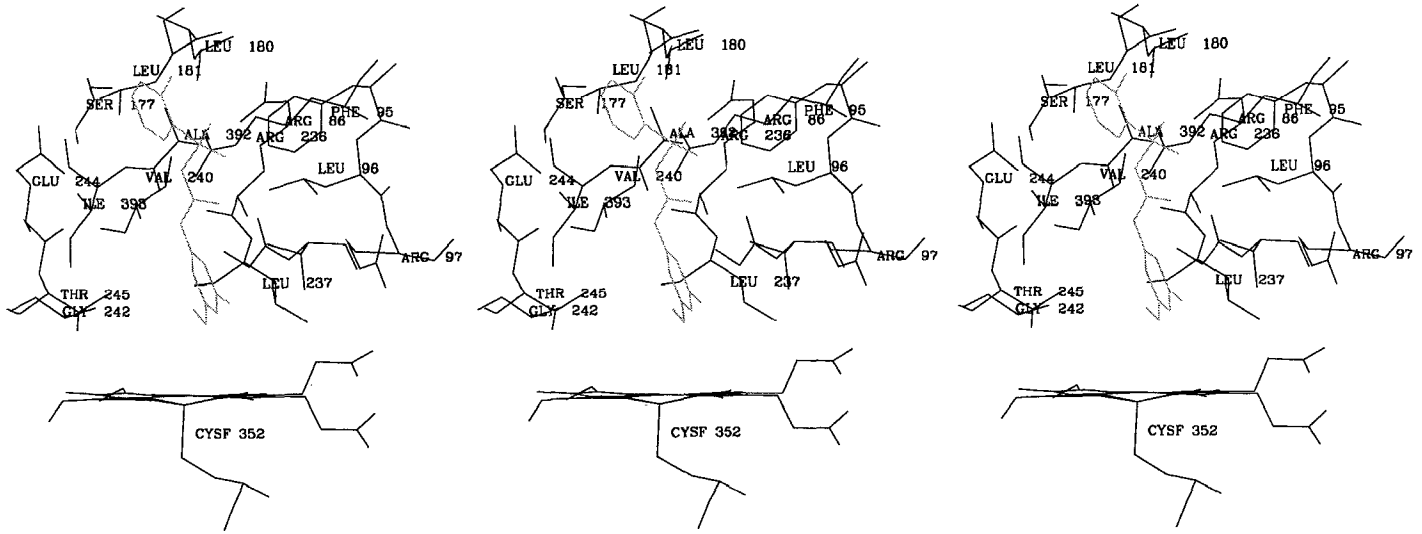

Fig. 4. Double stereoscopic image of the active site of the PILEUP model of SU2. This representation enables the reader to see both the front and back of the image. The substrate chlorsulfuron, shown in light grey lines, is docked in the active site. Arg^{86} and Arg^{236} are in good positions to provide hydrogen bonds to the sulfone group.

other charged or polar side chains in SU2 that are able to form hydrogen bonds with the substrate, but several hydrogen bonds can be formed between the urea part of the substrate and the backbone of the protein. Asp^{390} is also in the active site of the ISSC model; it is near the sulfone group and increases the amount of negative charge in that region. The unfavorable electrostatics might prevent binding of the substrate in the protein.

The three SU1 models are also fairly similar in the active-site region. Exceptions are Asp^{292}, which could interact with the nitrogen-substituted ring of chlorsulfuron, and Phe^{78} and Phe^{95}. In the ISSC model only Phe^{78} is in the active site, and in the second PILEUP model only Phe^{95} is in the active site. However, both occur in roughly the same region. This is because of the alignment of the B'-helix: both residues are at the C-terminal end of their respective B'-helices. These active sites have several hydrophobic groups near the substrate phenyl ring: Leu^{180}, Val^{181}, and the phenylalanines mentioned above. The polar residues that are in the active site have their side chains oriented away from the active site and thus cannot form hydrogen bonds to the substrate. Only the backbone atoms can form enzyme–substrate hydrogen bonds in these models. Furthermore, there are no positively charged side chains that are able to interact with the sulfone group in these models. Glu^{86} is near the sulfone group and the side chain could interact with the substrate near the sulfone group. However, since this is a negatively charged side chain, this interaction would be repulsive. Lys^{178} is the only positively charged residue in the region of the active site, and its side chain extends away from the active site.

Chlorsulfuron was able to fit into the ellipsoidally shaped SU1 and SU2 active sites, but not into the smaller spherically shaped P450cam active site. This is not surprising, considering the relative sizes and shapes of the molecules. Table 3 lists the active-site volumes of the models and P450cam. Since the models were not refined with the substrate in the active site, the docking of the substrate is not optimal. However, based on camphor in P450cam, the expected volume of the SU1 and SU2 active sites is approximately 710 $Å^3$. If the error of the estimate was 100 $Å^3$, the active-site volumes in Table 3 differ from the expected volume only by about 3.5 Å in radius, spread throughout the entire active site.

TABLE 3
ACTIVE-SITE VOLUMES OF THE MODELS

Model	Volume[a] ($Å^3$)
P450cam	510
SU1, ISSC, first minimization	640
SU1, ISSC, QPACK minimization	730
SU1, PILEUP (1), first minimization	950
SU1, PILEUP (1), qpack minimization	670
SU1, PILEUP (2), first minimization	660
SU1, PILEUP (2), qpack minimization	606
SU2, ISSC, first minimization	570
SU2, ISSC, QPACK minimization	650
SU2, PILEUP, first minimization	570
SU2, PILEUP, QPACK minimization	870

[a] Active-site volumes were determined by using a program which identifies internal void volumes. Using a grid of points (2.0 Å spacing) and the van der Waals radii of the atoms, spheres representing internal volumes were identified. The set of contiguous spheres in the active site was identified and the total volume was calculated. The volumes of camphor and chlorsulfuron are 290 and 405 $Å^3$, respectively.

SUMMARY AND CONCLUSIONS

Of the minimized models, two were chosen as the best representatives: the PILEUP models of SU1 and SU2. Both of these models conserved the heme cysteine, the threonine in the distal helix, and three of the five residues that interact with the heme propionates. These models also exhibit a patch of positive charge, similar to that of P450cam in the region believed to be the ferredoxin binding site, and their active sites are large enough to accommodate chlorsulfuron.

We found that the final structures obtained using different alignment methods were very similar for each protein. Figures 5 and 6 show the C^{α} traces of the models of SU1 and SU2, respectively. The overall topology of the models was conserved, leading to the conclusion that the final structure is relatively independent of the initial alignment. The most variable regions are the surface loops, which are not expected to affect the catalytic region significantly. However, the differences in alignment of the B'-helix in both SU1 and SU2 resulted in different residues in two positions in the active sites. Furthermore, an internal insertion in the SU1 sequence, that was modeled using two different backbone conformations from the PDB, caused an active-site residue, Phe^{95}, to adopt a substantially altered position in one of the models. The loop backbone conformations of four loops were the only differences between the two models. The early steps of energy minimization can also have a significant effect in regions of interest. Fixing the positions of large regions of the protein and relaxing them sequentially did not seem to make a significant difference; however, minimizing the residues with large van der Waals overlaps first resulted in large movements of residues in the region of the active site. This constituted about one-fourth of the total number of residues. For example, Arg^{97} substantially changes position in the QPACK-aided minimization of SU2, relative to the other energy minimization schemes employed. In the general approach, this residue interacts with the heme propionates to form a salt bridge. However, this is not a conserved salt bridge, and the side chain is in a distorted conformation. In the QPACK-aided refined structure, the side chain of this residue is approximately 180° away from the earlier conformation. Here the arginine extends to the molecular

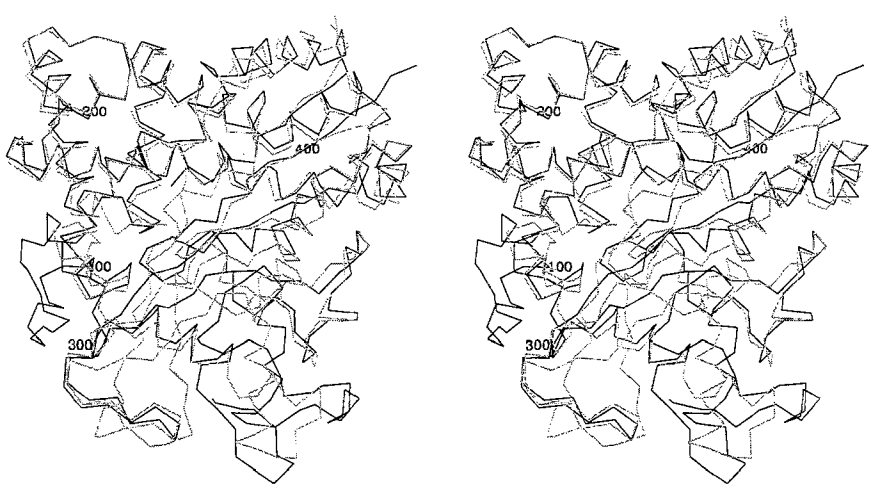

Fig. 5. Superimposed C^{α} traces of the two models of SU1. Dark and light grey lines represent the ISSC and PILEUP models, respectively. These models have better agreement in the helix-rich region (top of figure) than the SU2 models.

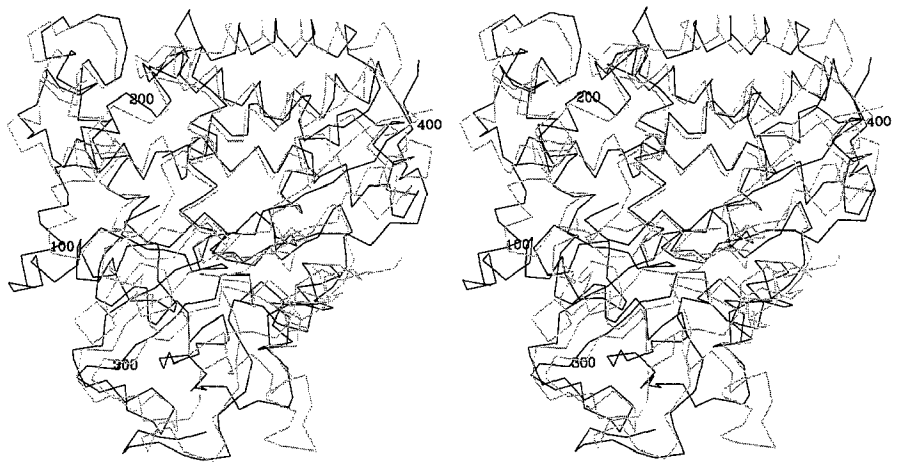

Fig. 6: Superimposed C^{α} traces of the two models of SU2. Dark and light grey lines represent the ISSC and PILEUP models, respectively. The B'-helix (at the left side of the figure, near residue 100) shows the greatest deviation of the helices; this helix contains active-site residues. The difference in the alignment of this region resulted in the appearance of different residues in the active site.

surface, and there is some rearrangement of the backbone atoms as well. Allowing only some of the residues, scattered throughout the protein, to minimize first enabled this strained residue to rearrange. However, the residues surrounding Arg^{97} do form part of the active site and are proposed to interact with the substrate. Substantial changes in the positions of these residues can influence the details of the interactions with the substrate, although both models suggest that these surrounding residues do interact with the substrate.

In this paper we have chosen to exclude insertions or deletions from occurring in the helices. Recent work with T4 lysozyme [28] suggests that insertions into an α-helix can be accommodated in two ways. First, the helical core is maintained and the helix is extended at the ends. This is somewhat similar to the method we chose to model SU1 and SU2. Second, the helix will undergo a local deformation to accommodate the residue. While a protein can fold into either of these conformations, computational methods will require significantly more time to evaluate which conformation will be adopted by a particular inserted residue or residues. Therefore, an early choice was made as to how insertions would be accommodated by the structure. If gaps were allowed in helices, some of the insertions could have occurred in the residues surrounding the active site. Therefore, some caution should be exercised in the interpretation of the details of these models.

Earlier studies, in which 3D models for members of the P450 superfamily have been proposed, have not addressed the extent to which their model depended on the methods used to create the model. Each chose a single method for sequence alignment and a single energy minimization technique. A potential problem with using the entire superfamily in multiple sequence alignment is that the template structure (P450cam) is the most different of the set. For this reason, we chose two proteins whose sequences show two of the highest number of identities to the P450cam sequence, and aligned only these with P450cam. A multiple alignment including another member with high identity to P450cam (P450lin) gave the same results (data not shown).

Since an accurate model of the active site of the enzyme is the object of any modeling exercise of this type, time should be taken to point out where the least confident regions of the enzyme

are. The region from the B'-helix to residue 100 is modeled, by far, the least accurately. This region contains an insertion into the P450cam sequence. The choice of a template for this region has the most dramatic effect on which residues are in the active site. In addition, choice of minimization techniques also dramatically alters which residues are in contact with the substrate. Additional experiments, such as site-directed mutagenesis, will have to be performed to identify which residues in this region are in the active site. The loop between the F and G helices also shows some variation between models. This, however, is not totally unexpected. In molecular dynamics simulations of P450cam [29–31], this region has high mobility and the average position of this loop can vary from simulation to simulation. Other regions, like the I helix, are highly conserved. Consequently, these regions vary little between alignments and refinement techniques. Of all the known P450 sequences, the SU1 and SU2 sequences have the highest sequence identity with P450cam. It is interesting that so much variability can occur in the active sites of structures modeled here. Other P450 sequences would be expected to have greater variability in the assignment of residues to the template structure.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Mark D. Paulsen, Dr. Richard J. Douthart, and Dr. John B. Nicholas (all of Pacific Northwest Laboratory) for helpful discussions. In addition, the authors thank one of the referees for extra diligent effort that resulted in finding several sequence errors, as noted in the paper.

Pacific Northwest Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RLO 1830. J.A.B. and M.B.B. had predoctoral and postdoctoral appointments administered by the Northwest College and University Association for Science, in affiliation with Washington State University under contract DE-AM06-76-RLO2225 with the U.S. Department of Energy, Office of Energy Research. This work was supported by the Laboratory Directed Research and Development Program of Pacific Northwest Laboratory (R.L.O.) and by a grant from the Health Effects and Life Science Research Division of the Office of Health and Environmental Research of the U.S. Department of Energy (R.L.O.).

REFERENCES

1 Black, S.D. and Coon, M.J., Adv. Enzymol. Relat. Areas Mol. Biol., 60 (1987) 35.
2 Guengerich, F.P., J. Biol. Chem., 266 (1991) 10019.
3 Poulos, T.L., Finzel, B.C. and Howard, A.J., J. Mol. Biol., 195 (1987) 687.
4 Raag, R. and Poulos, T.L., Biochemistry, 28 (1989) 917.
5 Furuya, H., Shimizu, T., Hirano, K., Hatano, M., Fujii-Kuriyama, Y., Raag, R. and Poulos, T.L., Biochemistry, 28 (1989) 6848.
6 Shimizu, T., Sadeque, A.J.M., Sadeque, G.N., Hatano, M. and Fujii-Kuriyama, Y., Biochemistry, 30 (1991) 1490.
7 Iwasaki, M., Juvonen, R., Lindberg, R. and Negishi, M., J. Biol. Chem., 266 (1991) 3380.
8 Graham-Lorence, S., Khalil, M.W., Lorence, M.C., Mendelson, C.R. and Simpson, E.R., J. Biol. Chem., 266 (1991) 11939.
9 Morris, G.M. and Richards, W.G., Biochem. Soc. Trans., 19 (1991) 793.
10 Laughton, C.A., Neidle, S., Zvelebil, M.J.J.M. and Sternberg, M.J.E., Biochem. Biophys. Res. Commun., 171 (1990) 1160.
11 Zvelebil, M.J.J.M., Wolf, C.R. and Sternberg, M.J.E., Protein Eng., 4 (1991) 271.

12 Nebert, D.W. and Gonzales, F.J., Annu. Rev. Biochem., 56 (1987) 945.
13 Omer, C.A., Lenstra, R., Litle, P.J., Dean, C., Tepperman, J.M., Leto, K.J., Romesser, J.A. and O'Keefe, D.P., J. Bacteriol., 172 (1990) 3335.
14 Harder, P.A., O'Keefe, D.P., Romesser, J.A., Leto, K.J. and Omer, C.A., Mol. Gen. Genet., 227 (1991) 238.
15 Unger, B.P., Gunsalus, I.C. and Sligar, S.G., J. Biol. Chem., 261 (1986) 1158.
16 Devereux, J., Haeberli, P. and Smithies, O., Nucleic Acids Res., 12 (1984) 387.
17 Rechid, R., Vingron, M. and Argos, P., CABIOS, 5 (1989) 107.
18 Dayhoff, M.O., Barker, W.C. and Hunt, H.T., Methods Enzymol., 91 (1983) 524.
19 Nelson, D.R. and Strobel, H.W., Mol. Biol. Evol., 4 (1987) 572.
20 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
21 Bass, M.B., Hopkins, D.F., Jaquysh, W.A.N. and Ornstein, R.L., Proteins, 11 (1992) 26.
22 Hagler, A.T., In Hruby, V.J. and Meienhofer, J. (Eds.) The Peptides, Vol. 7, Academic Press, New York, NY, 1985, pp. 213–299.
23 Dauber-Osguthorpe, P., Roberts, V.A., Osguthorpe, D.J., Wolff, J., Genest, M. and Hagler, A.T., Proteins, 4 (1988) 31.
24 Paulsen, M.D. and Ornstein, R.L., Proteins, 11 (1991) 184.
25 Arnold, G.E. and Ornstein, R.L., Proteins, 18 (1994) 19.
26 Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 211 (1990) 959.
27 Stayton, P.S. and Sligar, S.G., Biochemistry, 29 (1990) 7381.
28 Heinz, D.W., Baase, W.A., Dahlquist, F.W. and Matthews, B.W., Abstract S2, Sixth Symposium of the Protein Society, San Diego, CA, 1992.
29 Paulsen, M.D., Bass, M.B. and Ornstein, R.L., J. Biomol. Struct. Dyn., 9 (1991) 187.
30 Paulsen, M.D. and Ornstein, R.L., J. Comput.-Aided Mol. Design, 6 (1992) 449.
31 Paulsen, M.D. and Ornstein, R.L., Protein Eng., 6 (1993) 359.
32 Doolittle, R.F., Science, 214 (1981) 149.