Pergamon

Bioorganic & Medicinal Chemistry, Vol. 3, No. 4, pp. 411–428, 1995

0968-0896/95 $9.50 + .00

0968-0896(95)00031-3

Conformational Changes of Small Molecules Binding to Proteins†

Marc C. Nicklaus,* Shaomeng Wang, John S. Driscoll and George W. A. Milne

Laboratory of Medicinal Chemistry, Developmental Therapeutics Program, Division of Cancer Treatment, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, U.S.A.

Abstract—Flexible molecules change their conformation upon binding to a protein. This was shown by the analysis of small molecules whose structures have been determined by X-ray crystallography of both the pure compound and the compound bound to a protein. Thirty-three compounds present both in the Cambridge Structural Database and the Brookhaven Protein Data Bank were analyzed, and both were compared with the global energy minimum conformation calculated by the molecular mechanics program CHARMm. It was found that the conformation bound to the protein differs from that in the crystal structure and also from that of the global energy minimum, and the degree of deformation depends upon the number of freely rotatable bonds in the molecule. Analysis of the conformational energies of the flexible molecules showed that, for most of those compounds, both the crystal and the protein-bound conformations are energetically well above the global minimum, and, in many cases, not even in any local energy minimum. Semi-empirical calculations performed for a select number of structures, using both the AM1 and PM3 hamiltonians, confirmed these results. These findings are discussed as to their impact upon contemporary methods of drug design.

Introduction

Exactly one century ago, Emil Fischer introduced the 'lock and key' concept[1,2] of the interaction between enzymes and their ligands. In the more recent past however, increased understanding of the flexibility and internal motion of both ligands and enzymes has demanded modification of the idea that both the key and the lock are rigid. The 'induced fit' theory proposed[3] in 1958 introduced the aspect of molecular flexibility in enzyme–substrate interactions. The focus here was on conformational changes in the enzyme but more recent work has drawn attention to the possibility of conformational changes in ligands, and the notion of a 'rusting of the lock-and-key model' has been put forth.[4]

With the growth of interest in molecular modeling as a method of drug design, there have been several recent reports which touch upon this question. Ricketts and coworkers[5] have compared the structures of free molecules with their structures when bound to proteins and Rice and coworkers[6] have examined the conformations of some oligosaccharides in solution *vis-à-vis* their conformations when bound to enzymes. Changes in nucleotide conformation upon protein binding were studied by Moodie and Thornton[7] and a statistical comparison of free and protein-bound conformations of small molecule ligands was performed by Klebe.[8,9]

Numerous studies have dealt with binding energies, and conformational energies of specific ligands have been reported in a number of papers,[8–10] but less work seems to be available that analyzes conformational energy changes of ligands in a comprehensive manner. Andrews and coworkers[11] obtained an indirect indication as to the deformation energies of enzyme-bound ligands from a statistical analysis of the binding constants of 200 drugs.

All of these reports raise concern that deformation of ligands during protein binding may be a widespread phenomenon and in view of the significance of such a finding in drug design, we undertook a systematic study of the subject. The conformations present in the single crystal and those present in the enzyme-bound form were recorded and compared to each other and both were then compared to the modeled global minimum energy structure. Both geometrical and energetic aspects of the results of these comparisons are dealt with in this paper.

Only *conformational* (deformation) energies have been analyzed in this study, and no attempt has been made to consider entropic, solvation, and desolvation effects, which would have to be included if conclusions were to be drawn as to the *binding affinities* of small ligands from the results presented here.

Methods and Materials

Structures used

A total of 33 structures, present in both the Cambridge Structural Database[12] (CSD) and the Protein Data

†A preliminary account of this work has been presented in part in: Nicklaus, M. C.; Wang, S.; Milne, G. W. A.; Driscoll, J. S. *207th ACS National Meeting*, San Diego, CA, COMP 133A, 1994.

Bank[13] (PDB) and listed in Table 1 were found to meet the following conditions: (i) structurally and stereochemically identical fragments were described in both databases, (ii) the fragment in question is a small molecule, not a peptide, inorganic ion or bulk solvent molecule, (iii) the fragment is not covalently bound to the protein in the PDB file, (iv) the macromolecule in the PDB file is a protein, not a nucleic acid or other

Table 1. Compounds used in this study, and the database entries from which they were derived

No.	Compound	CAS RN	Rotors	GS[a]	GSN[b]	CSD Reference Code	PDB File
1	Adamantane	281-23-2	0	1.70	0.44	ADAMAN01	4cpp
2	β-L-arabinose	5328-37-0	0	3.17	0.16	ABINOS, ABINOS01	1bap
3	*tert*-Butanol	75-65-0	0	2.22	0.15	VATSAK(3)[c]	7rsa
4	Dimethyl sulfoxide (DMSO)	67-68-5	0	2.00	0.20	BAKLEE(2), DMETSO, DMETSO01	1r09, 6adh(2)
5	Guanine	73-40-5	0	1.90	0.44	GUANMH10	1ulb
6	Imidazole	288-32-4	0	1.75	0.29	IMAZOL01, IMAZOL02, IMAZOL04, IMAZOL06, IMAZOL10, IMAZOL13	1mbi, 4mba
7	Adenosine-5'-diphosphate (ADP)	58-64-0	6	4.15	0.36	KADPHD01, KADPHD02	1pfk(4), 4pfk(2)
8	Adenosine-5'-triphosphate (ATP)	56-65-5	8	4.83	0.33	ADENTP(2)	4at1(2), 7at1
9	Biotin	58-85-5	5	3.79	0.28	BIOTIN01, BIOTIN10	1stp
10	Chloramphenicol	56-75-7	7	3.95	0.29	CLMPCL01, CLMPCL02	1cla, 3cla,4cla
11	Citric acid	77-92-9	5	3.30	0.21	CITRAC10	1cts, 3cts, 5ldh, 8ldh(2)
12	Dimethylformamide	68-12-2	1	2.38	0.15	NAIDMF	6est(3),7est(5)
13	Fructose-6-phosphate	643-13-0	4	4.12	0.18	DEFLAB, FAVBOT	1fbp(2), 4pfk, 5fbp(3)
14	FK506	104987-11-3	12[d]	4.19	0.51	FINWEE, FINWEE10	1fkf
15	α-D-galactose	3646-73-9	1	2.38	0.36	ADGALA01, ADGALA10	8abp, 9abp
16	β-D-galactose	7296-64-2	1	2.38	0.36	BDGLOS01, BDGLOS10	8abp, 9abp
17	β-D-glucose	492-61-5	1	2.38	0.36	GLUCSE, GLUCSE01	3gbp
18	p-Hydroxybenzoic acid	99-96-7	1	2.37	0.34	PHBZAC, PHBZAC01	2phh
19	Isocitric acid	320-77-0	5	3.32	0.21	KHICIT, KHICIT01	5icd
20	D-Malic acid	636-61-3	3	2.97	0.19	LIHMAM	4csc
21	L-malic acid	97-67-6	3	2.97	0.19	AMHMAM, AMHMAN, BUHPAV(2), CAMALH	1csc, 3csc
22	Malonic acid	141-82-2	2	2.64	0.20	MALNAC	2at1(2), 7at1(2)
23	Maltose	69-79-4	4	4.06	0.27	MALTOT	1mbp
24	Methotrexate	59-05-2	10	5.40	0.34	DOJZAD, DOJZAD01	3dfr, 4dfr(2)
25	Oxamic acid	471-47-6	1	2.20	0.25	AMOXAM	1ldm(2)
26	2-Phospho-D-glyceric acid	3443-57-0	4	3.14	0.20	DEFHOL	7enl
27	3-Phosphoglyceric acid	820-11-1	4	3.38	0.19	BAXSIC	3pgk
28	N-(phosphonacetyl)-L-aspartate (PALA)	60342-56-5	7	3.91	0.21	COYDUP	8atc(2)
29	Pyridoxamine-5'-phosphate	529-96-4	4	3.32	0.30	PYRPOC	2aat
30	Pyruvic acid	127-17-3	1	2.20	0.25	PRUVAC	3ldh
31	D-Sorbitol (D-glucitol)	50-70-4	5	3.50	0.15	GLUCIT, GLUCIT01	4xia
32	Sucrose	57-50-1	5	3.47	0.35	SUCROS03, SUCROS04, SUCROS11	1rla
33	Xylitol	87-99-0	4	3.27	0.13	XYLTOL	5xia(2)

[a]Global Simple flexibility index.[15]
[b]Global Simple Normalized flexibility index.[15]
[c]A number in parentheses denotes multiple occurrence of the molecule in the same database entry.
[d]Eight non-ring rotors + four effective macrocyclic rotors (see text).

biopolymer. Where multiple entries were found for the same compound in either database, then all the structures were used in the analysis. Likewise, multiple molecules in the unit cell (in the CSD) and multiple occurrences of a ligand in different binding sites of a protein in the PDB were also included. In this way, the 33 original compounds provided 63 individual structures from the CSD and 72 from the PDB. These are listed in Table 1 together with the database entries from which they are derived. Details of the database search procedures used are given in the Experimental Section.

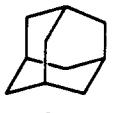
The precision of atomic coordinates in small molecule crystallography is almost always better than 0.01 Å and the systematic errors can be kept in the same range in diligent work.[14] Only substantially lower precision is usually achieved in macromolecule X-ray crystallography, where typical positional uncertainties reported are 0.2–0.3 Å. The resolution and *R*-factor achieved in the determination of the PDB structures used in this study are listed in Table 2. The average crystallographic resolution of all PDB structures containing flexible ligands was 2.25 ± 0.45 Å, the average *R*-factor 0.182 ± 0.028. These uncertainties are further examined in the Discussion.

Of the 33 compounds, the first six are rigid while the remaining 27 each possess at least one truly rotatable bond, or rotor, defined as a single bond to a group other than terminal CH_3, NH_2, OH, etc., whose rotation does not produce any new conformation of heavy atoms. Only the 27 flexible molecules were used in the energetic analysis since the rigid compounds' lack of rotatable bonds renders the determination of torsional deformation energies meaningless.

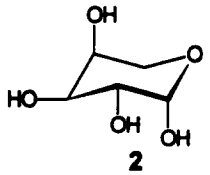
Table 2. Crystallographic parameters and proteins in the PDB files for the flexible compounds used in this study

No.	Compound	PDB File	Resolution (Å)	*R*-factor	Protein
7	ADP	1pfk	2.4	0.165	Phosphofructokinase
		4pfk	2.4	0.169	Phosphofructokinase
8	ATP	4at1	2.6	0.160	Aspartate carbamoyltransferase
		7at1	2.8	0.183	Aspartate carbamoyltransferase
9	Biotin	1stp	2.6	0.22	Streptavidin
10	Chloramphenicol	1cla	2.34	0.172	Chloramphenicol acetyltransferase
		3cla	1.75	0.157	Chloramphenicol acetyltransferase
		4cla	2.0	0.157	Chloramphenicol acetyltransferase
11	Citric acid	1cts	1.7	0.183	Citrate synthase
		3cts	1.7	0.192	Citrate synthase
		5ldh	2.7	0.196	Lactate dehydrogenase
		8ldh	2.8	0.191	Lactate dehydrogenase
12	Dimethylformamide	6est	1.8	0.2	Elastase
		7est	1.8	0.19	Elastase
13	Fructose-6-phosphate	1fbp	2.5	0.215	Fructose-1,6-biphosphatase
		4pfk	2.4	0.169	Phosphofructokinase
		5fbp	2.1	0.177	Fructose-1,6-biphosphatase
14	FK506	1fkf	1.7	0.170	FK506 binding protein
15	α-D-galactose	8abp	1.49	0.175	L-Arabinose-binding protein
		9abp	1.97	0.180	L-Arabinose-binding protein
16	β-D-galactose	8abp	1.49	0.175	L-Arabinose-binding protein
		9abp	1.97	0.180	L-Arabinose-binding protein
17	β-D-glucose	3gbp	2.4	0.158	Galactose-binding protein
18	p-Hydroxybenzoic acid	2phh	2.7	0.168	*p*-Hydroxybenzoate hydroxylase
19	Isocitric acid	5icd	2.5	0.176	Isocitrate dehydrogenase
20	D-Malic acid	4csc	1.9	0.188	Citrate synthase
21	L-malic acid	1csc	1.7	0.188	Citrate synthase
		3csc	1.9	0.177	Citrate synthase
22	Malonic acid	2at1	2.8	0.170	Aspartate carbamoyltransferase
		7at1	2.8	0.183	Aspartate carbamoyltransferase
23	Maltose	1mbp	2.3	0.25	D-Maltodextrin-binding protein
24	Methotrexate	3dfr	1.7	0.152	Dihydrofolate reductase
		4dfr	1.7	0.155	Dihydrofolate reductase
25	Oxamic acid	1ldm	2.1	0.173	Lactate dehydrogenase
26	2-Phospho-D-glyceric acid	7enl	2.2	0.169	Enolase
27	3-Phosphoglyceric acid	3pgk	2.5	N/G[a]	Phosphoglycerate kinase
28	PALA	8atc	2.5	0.165	Aspartate carbamoyltransferase
29	Pyridoxamine-5'-phosphate	2aat	2.8	0.22	Aspartate aminotransferase
30	Pyruvic acid	3ldh	3.0	N/G[a]	Lactate dehydrogenase
31	D-Sorbitol	4xia	2.3	0.156	D-Xylose isomerase
32	Sucrose	1rla	3.2	0.293	Rhinovirus serotype 1 coat protein
33	Xylitol	5xia	2.5	0.147	D-Xylose isomerase

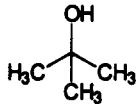
[a] *R*-factor not given.



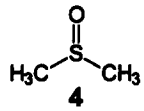
1



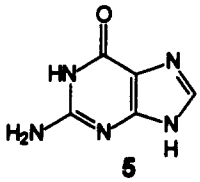
2



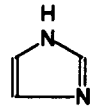
3



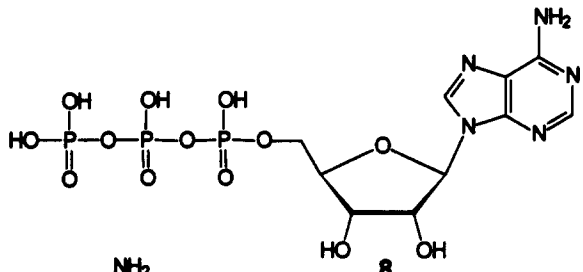
4



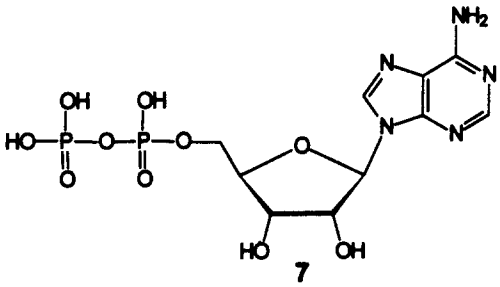
5



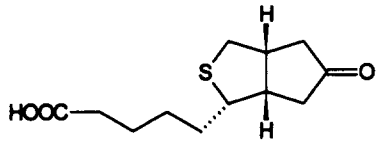
6



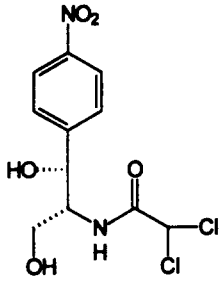
8



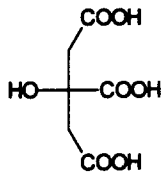
7



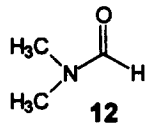
9



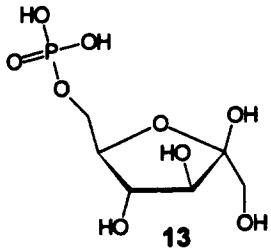
10



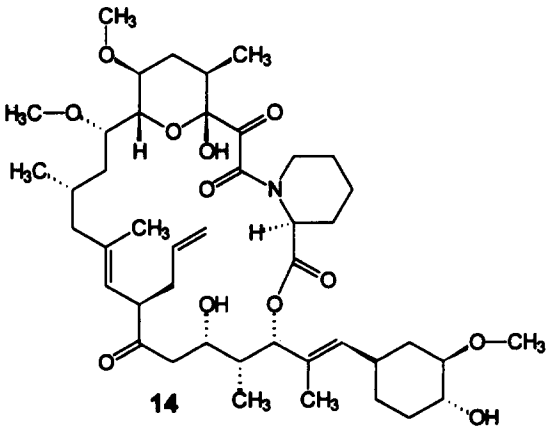
11



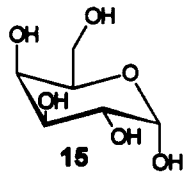
12

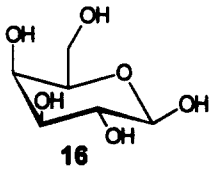


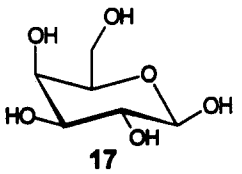
13

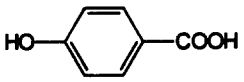


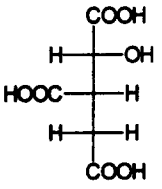
14

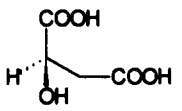

15

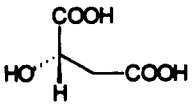

16

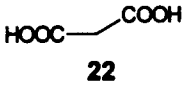

17

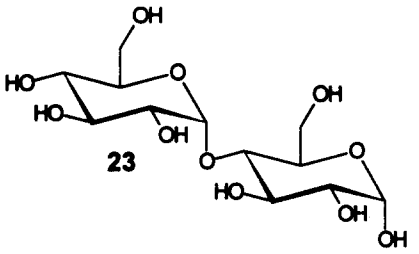

18

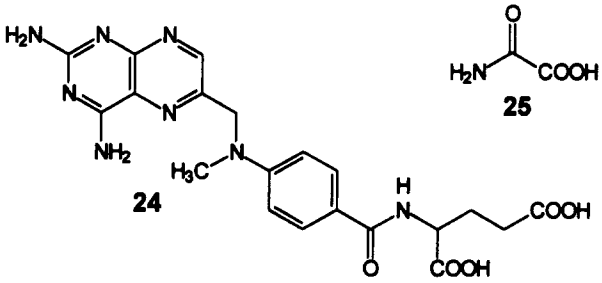

19


20

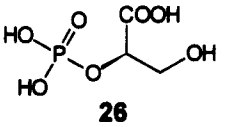

21

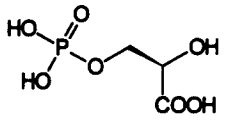

22

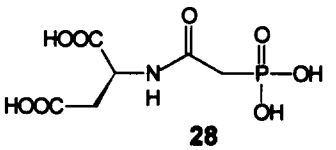

23

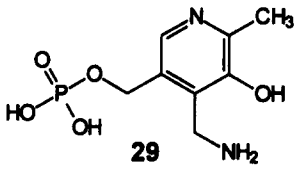

24

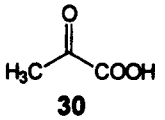
25

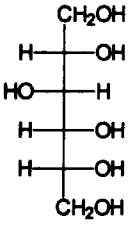

26

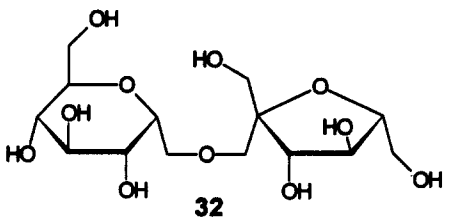

27

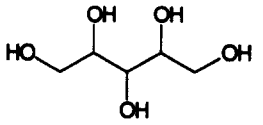

28


29


30


31


32


33

Flexibility measures

Three different measures of flexibility were used and are given in Table 1. The first of these is a simple count of the number of rotors in a structure. This is zero for the six rigid compounds and ranges from one to 10 for the remaining 27 compounds. Bonds in five and six-membered rings, not usually considered to be freely rotatable, were not counted. FK506 (**14**) contains a 21-membered ring. Five of the 21 macrocyclic bonds are either double bonds, torsions within six-membered rings, or amide bonds. Of the remaining ring bonds, only about 12 showed an appreciable variation in their dihedral angle in an extensive conformational study of the molecule.[10] Because of the constraints imposed by ring closure, a fractional rotor value of 1/3 was assigned to each of these bonds, yielding a total of 12 rotors (8 non-ring rotors + 12/3 ring rotors) for FK506.

The second measure of flexibility that was used was the Global Simple flexibility index (GS), developed by Fisanick and coworkers.[15] This is a function of the shortest topological paths between all pairs of (non-hydrogen) atoms in a structure and takes into account the types of bonds and the extent of branching in a path. The third measure, the Global Simple Normalized flexibility index (GSN) is a normalized version of the GS, with a scale of 1.0 to 0.0. The GSN is an inverted measure in that the more flexible a structure, the lower its GSN index.

Molecular modeling

Molecular mechanics modeling. Molecular mechanics modeling was used (i) to develop conformationally faithful structures of the 27 flexible molecules in the ligand–protein complexes from the PDB as well as in the single crystal unit cells from the CSD and (ii) to calculate the global minimum energy structure of each compound. All modeling was carried out using Version 3.3 of QUANTA,[16] run on a Silicon Graphics 4D/310GTX workstation operating under IRIX version 4.0.5.[17] Energy calculations were performed using version 22.0 of CHARMm[18] with the associated standard parameter set.[19] Non-bonded interactions were cut off at 14.0 Å, using a shift function to smooth the transition for both the van der Waals and electrostatic interactions. The computations used standard template-derived charges for which the CHARMm force field calculations were optimized. All non-bonded terms, including electrostatics, were included and a dielectric constant CDIE = 1 was used. All energy minimizations were carried out with an Adopted Basis Newton–Raphson algorithm,[18,20] which was allowed to iterate through 9990 steps or until convergence, defined as an energy gradient of $\leq$ 0.001 kcal mol^{-1} Å^{-1}, was achieved.

All structures were modeled in their un-ionized state. Both in the crystals and the protein–ligand complexes, a local ionic charge is always globally neutralized by some opposite charge in its surroundings, and it is quite often neutralized locally by a counterion. This is modeled more faithfully with neutralized acidic and basic groups[21] since the vacuum force field calculations would otherwise be dominated by artificially high electrostatic interactions.

In order to reliably determine the conformational energies of experimentally obtained structures, it is necessary to ascertain that differences between the measured bond lengths and the theoretical equilibrium bond lengths used in the modeling approaches do not lead to artifactually high energies. Ricketts *et al.*[5] have discussed bond length incompatibilities between various 3D coordinate sources such as experimental databases or modeling programs. These bond length differences were eliminated in our study as follows. The experimental structures—both from the CSD and the PDB—were examined to determine all the torsion angles in the ligand and then the ligand molecule was modeled independently as described above, using these torsions, which were held constant during energy minimization of the structure. Bond angles occupy a middle ground between bond lengths and torsion angles. Deviations from the equilibrium bond angles used in CHARMm which clearly represented conformational changes (bond angles between large functional groups) were modeled in the derived structures in an analogous manner as for the torsion angles. The remaining bond angles were allowed to relax freely. Compounds containing five- or six-membered non-aromatic rings were modeled with a local minimum ring conformation that represented the experimental structure most faithfully. The 21-membered ring structure FK506 (**14**) was treated in a slightly different fashion detailed in the Experimental Section. Typical RMS deviations (for all heavy atoms) between the atomic coordinates of derived and the database-stored structures were in the range of 0.1–0.4 Å. A few cases in which we experienced greater than usual difficulties in obtaining an acceptable RMS deviation *and* simultaneously a realistic conformational energy are likely due to a questionable validity of the experimental data, and are listed in the Experimental Section.

All but one PDB and many CSD structures lacked hydrogen coordinates, and for those structures, the construction of the derived models in QUANTA was always followed by an energy minimization of the model's hydrogens while keeping all heavy atoms fixed. For those experimental structures that *did* include hydrogen coordinates, these hydrogen positions were used in the derived structure where a torsion was involved, i.e. typically for hydroxyl or methyl hydrogens, but not, e.g., for the hydrogens on rigid rings. In order to obtain conformational energies that could be compared with those of the hydrogen-less experimental structures, the conformational energies were also calculated after energy-minimizing the hydrogens in exactly the same way as described above.

Conformational searches were used for all non-rigid compounds to determine the global energy minimum structure. For all rotatable bonds, a conformational search was performed using the Grid Search capability within QUANTA to identify the most stable conformer in vacuum. Because such searches can be very time-consuming, the step size was set to 120° for all simple sp^3–sp^3 bonds and 90°, 60° or 30° for bonds where more intramolecular interaction was present. With a few compounds, such as **7**, **8** and **10**, which contain many rotors, a random sampling procedure was used instead of an exhaustive grid search. In this procedure, the average number of angle steps was approximately two. In all conformational searches, every conformation explored was allowed to relax fully from its starting point on the multi-torsion grid, which provides the highest assurance that all low-energy local minima are found, including, hopefully, the global minimum. As a check of this procedure, a very exhaustive random search was carried out for pyridoxamine-5'-phosphate (**29**), exploring 18,000 conformations, and the global energy minimum so obtained was identical to that obtained with the much less exhaustive 'standard' grid search which had investigated 864 conformations. For all compounds containing five- or six-membered non-aromatic rings, e.g. **2**, **7**, **8** and **9**, the conformational searches were repeated using the different local minimum energy ring conformations (such as boat/chair) in the starting conformer. Torsions that were not counted as true rotors in the *experimental* structures, such as hydroxyl torsions, were also included in the conformational searches in many instances because they are relevant for the energy of the *modeled* structures, which do carry hydrogens.

Coordinates of all modeled structures as well as the detailed strategies used in the individual conformational searches are available from the authors upon request.

Semi-empirical modeling. In order to compare the molecular mechanics calculations with a higher-level method, a limited number of semi-empirical energy computations were performed. Both the AM1[22] and the PM3[23] hamiltonians were used in MOPAC 6.0,[24] and in all cases the keyword 'PRECISE' was employed. The hydrogens necessary for the MOPAC calculations were added to the experimental structures in QUANTA, and the resulting structures were then imported into MOPAC, where the hydrogens were freely re-minimized. Bond lengths and angles were allowed to relax to remove incompatibilities between the experimental values and the values derived from either hamiltonian. No bond angle re-adjustments analogous to the ones applied to the CHARMm models were performed in MOPAC. It is also not computationally feasible to perform exhaustive conformational searches using semi-empirical methods, and so the QUANTA-calculated global energy minima were used after full re-minimization in MOPAC.

Geometric structure comparisons

For each of the 33 compounds in this study, the protein-bound conformation was compared both with the structure given by the single crystal X-ray diffraction study and with the global minimum structure calculated by QUANTA/CHARMm. For those compounds for which multiple database entries and/or multiple occurrences of the molecule in the same entry were found in either database, the average of all possible PDB–CSD comparisons was calculated. A similar averaging was performed when comparing the protein-bound structure(s) to the global energy minimum conformation. For the rigid compounds, the PDB coordinates were compared with the coordinates of a model built with QUANTA. The global energy minimum conformation obtained by conformational searching of the flexible molecules were compared with the derived models of the protein-bound structures since these were the structures used in the subsequent energy comparisons.

Comparison between two structures was carried out using the Molecular Similarity functionality provided as part of QUANTA. A Rigid Body Fit was performed on all non-hydrogen ('heavy') atoms, and the result is recorded in Table 3 as the RMS value, the sum of the root mean square of all the heavy atom displacements.

Conformational energy comparisons

For each of the 27 flexible molecules, the CHARMm energy was computed for each derived protein-bound and single-crystal conformation. The difference between these energies and the global energy minimum is listed in Table 4 as 'ΔE_{global}' and is henceforth also referred to as 'global energy'.

In order to analyze how far above the nearest local energy minimum each of the experimental conformations was located, each one of the derived experimental structures was allowed to relax fully by the minimization procedure described above. The global minimum energy was then subtracted from the CHARMm energy of this local energy minimum, and the resulting difference is reported in Table 4 as 'ΔE_{local}', and is called 'local energy' in the rest of this paper.

Both the global and the local energies were calculated for experimental structures with energy-minimized hydrogen positions, designated in Table 4 as 'H's minim.'. For those structures that carried hydrogens in the database entry, global and local energies were also calculated using the experimental hydrogen coordinates before minimization, and these conformational energies are identified in Table 4 by the subheading 'H's unmin.'.

Table 4 also lists the number of atoms in each ligand that are capable of forming hydrogen bonds (H-bonds),

Table 3. Average RMS deviation between conformations in PDB and CSD (RMS_{CSD}) and between conformations in PDB and calculated global energy minimum structure ($RMS_{Glob.Min.}$)

No.	Compound	RMS_{CSD} (Å)	$RMS_{Glob.Min.}$ (Å)
Rigid Compounds			
1	Adamantane	0.08	0.048
2	β-L-Arabinose	0.043 ± 0.002	0.079
3	*tert*-Butanol	0.068 ± 0.015	0.040
4	DMSO	0.026 ± 0.011	N/C[a]
5	Guanine	0.04	0.059
6	Imidazole	0.025 ± 0.002	0.035 ± 0.005
Flexible Compounds			
7	ADP	1.62 ± 0.41	2.72 ± 0.47
8	ATP	2.63 ± 0.31	3.49 ± 0.09
9	Biotin	1.18	2.20
10	Chloramphenicol	0.584 ± 0.032	1.27
11	Citric acid	1.04 ± 0.25	1.50 ± 0.11
12	Dimethylformamide	0.394 ± 0.033	0.378 ± 0.015
13	Fructose-6-phosphate	0.41 ± 0.21	1.79 ± 0.09
14	FK506	2.96	3.92
15	α-D-Galactose	0.485 ± 0.007	0.112 ± 0.014
16	β-D-Galactose	0.129 ± 0.002	0.080 ± 0.011
17	β-D-Glucose	0.554 ± 0.001	0.124
18	*p*-Hydroxybenzoic acid	0.063 ± 0.003	0.012
19	Isocitric acid	0.716 ± 0.050	1.74
20	D-Malic acid	0.294	1.19
21	L-Malic acid	0.84 ± 0.36	0.90 ± 0.51
22	Malonic acid	0.501 ± 0.087	0.91 ± 0.20
23	Maltose	0.579	0.366
24	Methotrexate	2.68 ± 0.16	3.41 ± 0.10
25	Oxamic acid	0.188 ± 0.034	0.093 ± 0.007
26	2-Phospho-D-glyceric acid	1.09	1.56
27	3-Phosphoglyceric acid	0.858	1.59
28	PALA	1.46 ± 0.18	0.84 ± 0.28
29	Pyridoxamine-5-phosphate	0.569	1.00
30	Pyruvic acid	0.141	0.017
31	D-Sorbitol	1.01	1.93
32	Sucrose	1.72	1.64
33	Xylitol	0.915 ± 0.039	1.110 ± 0.085

[a] Not computable in CHARMm (incomplete parametrization).

which are presumably[25] a major contributor of binding energy in many protein–ligand complexes. These 'hydrogen bond centers' were defined as the elements N, O, and S with their usual hydrogen donor and/or acceptor capabilities. Table 4 also reports a simple H-bond count for each protein–ligand complex. This number, which includes a certain uncertainty because of the low resolution of macromolecule X-ray structures, was determined by counting all non-covalent intermolecular bonds originating from any ligand H-bond center, with a donor–acceptor distance of up to 3.4 Å. H-bonds to water molecules were included, as well as non-covalent interactions mediated by metal ions such as Mg^{2+}. Six of the compounds (**7**, **19**, **26**, **27**, **31**, **33**) showed the presence of counterions in the protein–ligand complexes.

Results

Geometric changes

The results of the geometric comparisons of the protein-bound conformations with the single-crystal conformations are reported in Table 3 as RMS_{CSD}. For the six rigid compounds (**1–6**), the RMS_{CSD} value is small, ranging from 0.02 to 0.08 Å, with an average value of 0.047 ± 0.023 Å. Among the non-rigid structures **7–33**, the value of RMS_{CSD} varies from 0.188 to 2.96 Å, evidently increasing as the number of rotors in the structure increases. A plot of RMS_{CSD} versus the number of rotors in the structure is shown in Figure 1. There is a fair correlation (r^2 = 0.824) between these two quantities. The slope of the line in Figure 1 suggests that each rotor contributes about 0.24 Å to the value of RMS_{CSD}.

When either the GS or the GSN flexibility index was used to replace the number of rotors as an index of structural flexibility, the correlation with RMS_{CSD} was poorer. With GS, an r^2 of 0.642 was obtained and with GSN, r^2 = 0.090. The simple rotor count was therefore a better indicator of flexibility and was used exclusively in this study. This finding is in line with conclusions recently reached by Clark *et al.*[26]

Table 4. Conformational energies of the crystal and protein-bound structures

No.	Compound	No. of H-bond centers	Experimental structures[a]	No. of H-bonds counted[b]	ΔE_{local}[c]		ΔE_{global}[d]	
					H's minim.	H's unmin.[e]	H's minim.	H's unmin.[e]
7	ADP	14	KADPHD01		10.90		19.80	
			KADPHD02		5.80	8.66	25.97	28.83
			1pfk (4)[f]	20	5.91		20.08	
				17	8.33		22.11	
				15	3.96		21.66	
				16	9.77		22.31	
			4pfk (2)	12	5.11		21.38	
				16	8.67		21.21	
8	ATP	17	ADENTP (2)		18.74		29.31	
					19.23		32.02	
			4at1 (2)	10	15.34		34.12	
				9	22.84		35.96	
			7at1	7	18.85		41.63	
9	Biotin	6	BIOTIN01		0.52		4.47	
			BIOTIN10		0.52		4.47	
			1stp	12	5.52		8.22	
10	Chloramphenicol	6	CLMPCL01		4.87	8.58	12.32	16.03
			CLMPCL02		5.34	9.52	7.22	11.40
			1cla	7	7.80		19.10	
			3cla	9	8.12		19.42	
			4cla	5	9.31		20.62	
11	Citric acid	7	CITRAC10		4.69	4.83	22.95	23.09
			1cts	13	5.51		16.39	
			3cts	9	7.73		18.60	
			5ldh	4	8.92		12.04	
			8ldh (2)	6	22.08		36.86	
				7	14.33		33.44	
12	Dimethylformamide	2	NAIDMF		0.00		0.00	
			6est (3)	2	*9.90*[g]		*9.90*	
				2	*9.53*		*9.53*	
				1	*9.75*		*9.75*	
			7est (5)	1	*9.84*		*9.84*	
				1	*9.82*		*9.82*	
				1	*9.87*		*9.87*	
				1	*9.91*		*9.91*	
				3	*9.37*		*9.37*	
13	Fructose-6-phosphate	9	DEFLAB		9.23	18.82	31.72	41.31
			FAVBOT		11.52	22.15	30.90	41.53
			1fbp (2)	8	6.76		16.11	
				8	6.66		16.02	
			4pfk	14	7.21		16.56	
			5fbp (3)	13	12.96		24.60	
				20	3.29		16.27	
				13	4.51		15.79	
14	FK506	13	FINWEE		30.04	33.75	39.67	43.38
			1fkf	9	14.13		36.53	
15	α-D-galactose	6	ADGALA01		2.53	9.66	9.43	16.56
			ADGALA10		1.01	4.43	22.62	26.05
			8abp	9	1.39		1.39	
			9abp	8	1.05		1.05	
16	β-D-galactose	6	BDGLOS01		2.16	21.78	2.18	21.79
			BDGLOS10		1.78	19.30	1.80	19.32
			8abp	9	0.51		0.51	
			9abp	7	0.73		0.73	
17	β-D-glucose	6	GLUCSE		9.79	27.94	10.96	29.10
			GLUCSE01		2.69	24.53	3.86	25.70
			3gbp	13	1.81	15.09	4.98	18.26
18	*p*-Hydroxybenzoic acid	3	PHBZAC		-0.06[h]	0.02	0.06	0.14
			PHBZAC01		0.12	0.12	0.14	0.14
			2phh	7	0.00		0.00	

Table 4. *Continued.*

19	Isocitric acid	7	KHICIT		13.84	38.02	22.07	46.24
			KHICIT01		13.73	37.03	21.95	45.26
			5icd	16	7.55		21.35	
20	D-Malic acid	5	LIHMAM		1.20		12.11	
			4csc	11	3.20		14.10	
21	L-malic acid	5	AMHMAM		13.92	16.40	13.92	16.40
			AMHMAN		6.31	10.34	18.78	22.81
			BUHPAV (2)		0.72	13.09	14.41	26.78
					1.03	17.87	11.93	28.78
			CAMALH		5.65	22.55	16.55	33.45
			1csc	10	3.28		14.18	
			3csc	10	6.83		6.83	
22	Malonic acid	4	MALNAC		0.10		0.10	
			2at1 (2)	5	0.63		0.63	
				4	0.52		0.52	
			7at1 (2)	5	0.41		0.41	
				5	0.78		0.78	
23	Maltose	11	MALTOT		5.00	15.21	20.68	30.88
			1mbp	17	5.66		5.66	
24	Methotrexate	12	DOJZAD		23.00		29.50	
			DOJZAD01		9.46		28.25	
			3dfr	16	25.67		29.48	
			4dfr (2)	13	14.27		23.83	
				13	18.72		28.50	
25	Oxamic acid	4	AMOXAM		0.10		0.10	
			1ldm (2)	6	0.46		0.46	
				2	0.37		0.37	
26	2-Phospho-D-glyceric acid	7	DEFHOL		10.24	15.51	17.94	23.20
			7enl	9	21.39		36.29	
27	3-Phosphoglyceric acid	7	BAXSIC		18.18	18.62	18.18	18.62
			3pgk	6	19.86		22.08	
28	PALA	9	COYDUP		12.95		25.47	
			8atc (2)	18	11.71		14.23	
				17	16.56		16.77	
29	Pyridoxamine-5'-phosphate	7	PYRPOC		1.45		5.50	
			2aat	8	6.72		13.16	
30	Pyruvic acid	3	PRUVAC		0.02	0.09	2.66	2.72
			3ldh	3	0.02		0.02	
31	D-Sorbitol	6	GLUCIT01		4.00	16.02	21.65	33.67
			4xia	13	6.09		17.87	
32	Sucrose	11	SUCROS11		22.20	32.52	28.92	39.24
			1r1a	4	*29.62*		*57.37*	
33	Xylitol	5	XYLTOL		11.28	22.15	15.47	26.34
			5xia (2)	13	3.00		11.85	
				12	1.69		10.54	

[a]Uppercase: CSD structures; lowercase: PDB structures.
[b]Hydrogen bonds between the ligand and the protein in the PDB complexes.
[c]CHARMm energy difference to nearest local energy minimum (kcal mol^{-1}).
[d]CHARMm energy difference to global energy minimum (kcal mol^{-1}).
[e]Energy calculated with unminimized experimental hydrogen positions (see text); column left empty when no hydrogens present in the database entry.
[f]A number in parentheses denotes multiple occurrence of the molecule in the unit cell (for CSD structures) or in the macromolecule–ligand complex (for PDB structures).
[g]Italicized energy values indicate that the validity of the experimental data is in doubt; see text.
[h]The narrow global energy minimum was not caught by the conformational search procedure.

The results of the comparisons of the protein-bound conformations with the calculated global energy minimum conformation are listed in Table 3 as $RMS_{Glob.Min.}$. For the rigid compounds, the $RMS_{Glob.Min.}$ value again is small, ranging from 0.035 to 0.079 Å, with an average value of 0.052 ± 0.018 Å for **1–3**, **5**, **6**. For the flexible structures **7–33**, the value of $RMS_{Glob.Min.}$ varies from 0.017 to 3.92 Å, hence encompassing a larger range than RMS_{CSD}. Figure 2 shows $RMS_{Glob.Min.}$ plotted vs the number of ligand rotors. Again, a fair correlation (r^2 = 0.827) was obtained suggesting that these two quantities are not unrelated. While the strength of the correlation was essentially the same as for RMS_{CSD} vs number of rotors, the typical *amount* of RMS deviation observed between the protein-bound conformations and the global energy

minimum conformation was substantially larger than the average RMS_{CSD} per rotor. From Figure 2 it appears that each rotor contributes about 0.33 Å to the value of $RMS_{Glob.Min.}$.

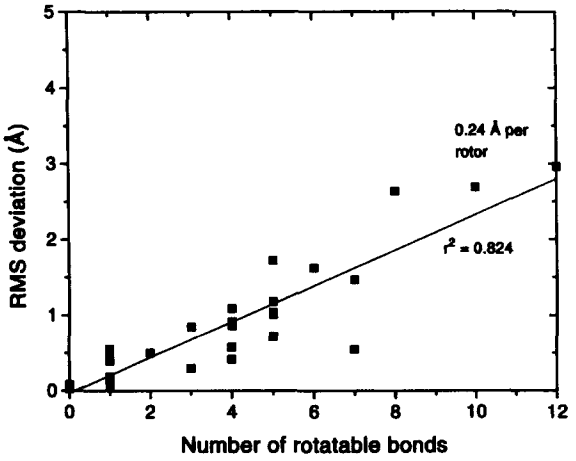


Figure 1. RMS differences between protein-bound and single crystal conformations of compounds **1–33**. Also shown is the regression line $RMS_{CSD} = 0.235(\pm 0.020)\ N_{Rot} - 0.030(\pm 0.090)$ ($n = 33$, SD = 0.337, $r^2 = 0.824$).

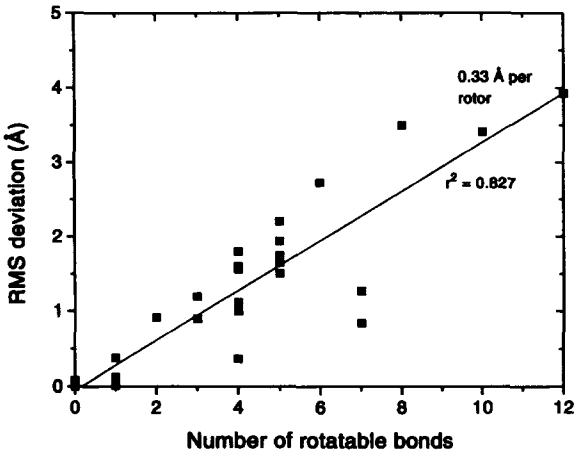


Figure 2. RMS differences between protein-bound and global energy minimum conformations (CHARMm vacuum energies) for compounds **1–33**. Estimated value of 0.0 Å used for **4** (see Table 3). Also shown is the regression line $RMS_{Glob.Min.} = 0.332(\pm 0.027)\ N_{Rot} - 0.050(\pm 0.125)$ ($n = 33$, SD = 0.470, $r^2 = 0.827$).

When GS and GSN were used instead of the number of rotatable bonds, essentially the same trends emerged for $RMS_{Glob.Min.}$ as had been found for RMS_{CSD}. A somewhat lower correlation of $RMS_{Glob.Min.}$ with GS ($r^2 = 0.674$) was obtained than with the number of rotatable bonds, whereas GSN, again, did not produce any discernible correlation ($r^2 = 0.043$).

Energetic changes

Table 4 reports the conformational energies of both the single-crystal and the protein-bound conformations of the 27 flexible compounds (**7–33**). The values listed are the differences between the CHARMm energies of the derived experimental structures and the global minimum conformation ('ΔE_{global}') or the nearest local minimum ('ΔE_{local}'). Conformational energies are generally given for structures with energy-minimized hydrogen positions ('H's minim.'). Since this conformation is a local minimum (in vacuum) with respect to the hydrogens, the actual hydrogen positions in the binding site might correspond to a higher conformational vacuum energy of the whole molecule. For those experimental structures that contained hydrogen coordinates in the database entry, conformational energies were also calculated using these hydrogen positions ('H's unmin.'). Because of the uncertainty of the hydrogen positions even in small molecule X-ray crystallography, the calculated energy of this conformation could conceivably be higher than the energy of the 'true' conformation in the crystal.

Energies well above the global minimum were found for many of the single crystal (CSD) conformations. In a number of instances (e.g. **13**, **28**, **33**), those energies were substantially higher than the energies of the same molecule bound to the protein. However, this was not a systematic trend since for about an equal number of compounds (e.g. **9**, **10**, **26**), the PDB energies were higher than the CSD energies. For the majority of the compounds, the conformational energies of the unbound and bound structures are close to each other.[7]

The conformational energies of the protein-bound structures showed that most of the ligands are not found in their vacuum global energy minimum conformation when lodged in the binding site, and energies well above the global minimum were found in many cases. The average global conformational energy for all protein-bound ligands was 15.9 ± 11.5 kcal mol^{-1}. A plot of these energies *vs* the number of ligand rotors (Fig. 3) shows that the typical amount of deformation energy increases with the flexibility of the molecules. A fair correlation ($r^2 = 0.633$) was obtained between these two quantities. The slope of the regression line shows that each rotor accounts for an average deformation energy of 3.4 kcal mol^{-1}. The global conformational energies are also increase ($r^2 = 0.493$) with the number of atoms in each ligand that are capable of forming H-bonds (Fig. 4), which supports the notion[25,27] that, in many cases, H-bonds provide a substantial part of the binding energy. Figure 4 suggests that the conformational deformation energy increases, on average, by about 2.0 kcal mol^{-1} for each additional H-bond center.

Inspection of Table 4 shows that many of the ligands are not even in any *local* vacuum energy minimum when complexed with the protein. The average local energy was 7.9 ± 6.8 kcal mol^{-1}. The local energy was typically slightly larger than half (0.60 ± 0.29; $r^2 = 0.727$) the global energy, which means that the remaining deformation energy after minimization ($\Delta E_{global} - \Delta E_{local}$) was on average *ca* 40 % of the energy of the unminimized structure. The local energy, i.e. the distance in energy from the nearest local minimum,

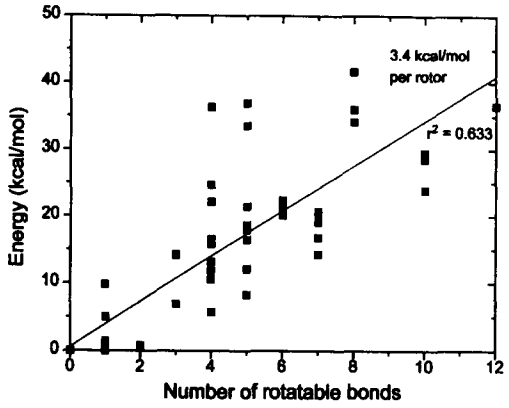


Figure 3. Conformational energies of the flexible compounds (**6–33**) vs number of ligand rotors. Shown are the differences in energy (CHARMm) between the protein-bound and global energy minimum conformations ('global energies'). All individual structures are included except for **12**, **18**, and **32** (see text). Also shown is the regression line $\Delta E_{global} = 3.367(\pm 0.352)\ N_{Rot} + 0.550(\pm 1.863)$ (n = 55, SD = 7.00, r^2 = 0.633).

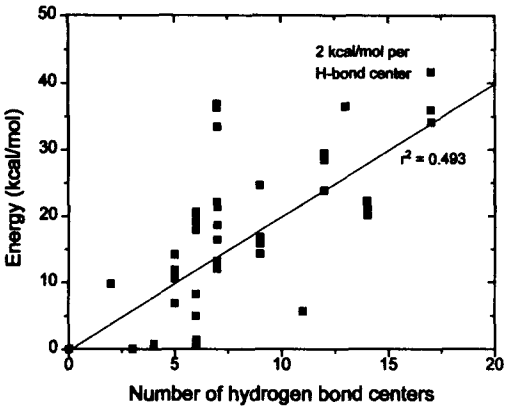


Figure 4. Conformational energies of the flexible compounds (**6–33**) vs number of ligand hydrogen bond centers. Shown are the differences in energy (CHARMm) between the protein-bound and global energy minimum conformations ('global energies'). All individual structures are included except for **12**, **18**, and **32** (see text). Also shown is the regression line $\Delta E_{global} = 2.010(\pm 0.280)\ N_{HbdCr} - 0.307(\pm 2.520)$ (n = 55, SD = 8.23, r^2 = 0.493).

exhibited a correlation with the number of ligand rotors of r^2 = 0.505 (Fig. 5). An average local energy per rotatable bond of 1.8 kcal mol^{-1} was obtained from the regression. A correlation of the local energy with the number of H-bond centers was present at the $P < 0.001$ level but weak (r^2 = 0.299; 0.9 kcal mol^{-1} per H-bond center).

No correlation (r^2 = 0.086) was found between the number of H-bonds counted in the protein–ligand complexes and the global energies. All energy values were however compatible with the number of H-bonds in the sense that none exceeded the total amount of H-bond energy that would be available if each H-bond formed had an energy of about 5 kcal mol^{-1}, the value generally accepted as the upper limit for geometrically optimal H-bonds of most types.[28] The average number of H-bonds formed per ligand H-bond center was 1.37 ± 0.54. This is in agreement with the numbers reported by

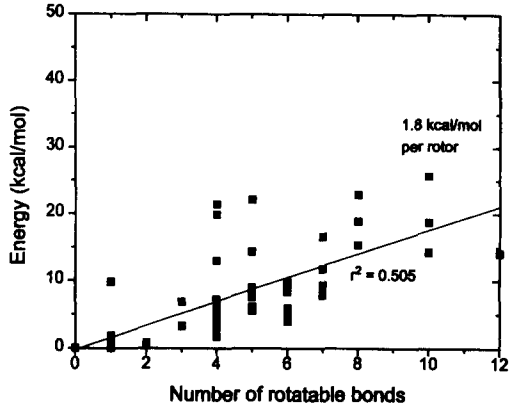


Figure 5. 'Local energies' for the flexible compounds (**6–33**), i.e. (CHARMm) energy differences between the protein-bound and the nearest local vacuum energy minimum conformation. All individual structures are included except for **12**, **18**, and **32** (see text). Also shown is the regression line $\Delta E_{local} = 1.779(\pm 0.242)\ N_{Rot} - 0.225(\pm 1.280)$ (n = 55, SD = 4.81, r^2 = 0.505).

Chakrabarti in a study of sulfate and phosphate ions bound to proteins,[29] who found an average number of 1.75 ± 0.75 H-bonds per oxygen if H-bonds to water molecules are included, and 1.25 ± 0.75 if only the H-bonds to the protein are counted.

Dividing the average global energy per H-bond center by the average number of H-bonds per H-bond center yielded an average global energy per H-bond formed that was on the order of 1.6 kcal mol^{-1}; the average of the ratios was 1.75 ± 1.57 kcal mol^{-1} per H-bond. These numbers are in line with an average value of 1.8 kcal mol^{-1} per H-bond found for sugars bound to proteins,[30] and also with a range of 0.24–1.7 kcal mol^{-1} found for the amide–amide H-bond strengths in a study[31] of binding constants of organic molecules.

Table 5 lists the semi-empirical AM1 and PM3 energies calculated with MOPAC 6.0 for a select set of compounds. These energies correspond to the *global* conformational energies (with minimized hydrogens) calculated by CHARMm (Table 4), i.e. they are the energy differences between the derived experimental structure and the global energy minimum. The sequence in which the results are given for multiple instances of a molecules in a PDB entry is the same as in Table 4. In most cases, the MOPAC energy of the QUANTA-calculated global energy minimum structure (after full re-minimization in MOPAC) was lower than that of any of the corresponding experimental conformations. In a very few cases, the MOPAC energy of a CSD structure (not reported) turned out to be slightly lower (by less than 3 kcal mol^{-1}) than that of the structure imported from QUANTA, and then, the lower energy structure was taken as the reference for calculating conformational deformation energies. The possibility of the existence of structures with still lower energies for both the AM1 and PM3 hamiltonians cannot be ruled out, but this question cannot be decided at present since it is not computationally feasible to perform

Table 5. AM1 and PM3 global conformational energies of select protein-bound structures

No.	Compound	Experimental structures	ΔE_{AM1} (kcal/mol)	ΔE_{PM3} (kcal/mol)
10	Chloramphenicol	1cla	8.30	5.22
		3cla	11.95	6.77
		4cla	13.61	6.94
12	Dimethylformamide	6est[a]	14.39	N/D[b]
15	α-D-galactose	8abp	3.03	3.57
		9abp	2.82	1.13
24	Methotrexate	3dfr	12.66	18.14
		4dfr (2)	18.03	23.10
			20.11	24.51
33	Xylitol	5xia (2)	4.53	4.40
			0.00	0.00

[a]Calculation performed only for the first of the three occurrences of DMF in 6est.
[b]Not determined.

exhaustive conformational searches using semi-empirical methods. Because of this, and because all bond angles were allowed to relax freely in the derived experimental structures in MOPAC, the possibility exists of a slight bias of the MOPAC results towards energy differences (global energies) that are too low for some of the structures. Keeping this in mind, the MOPAC energies agree well in a semi-quantitative manner with the molecular mechanics energies calculated by CHARMm. In particular, the high conformational energies of chloramphenicol (**10**) and methotrexate (**24**) are clearly confirmed.

Discussion

We have evaluated in this study a good part of the body of experimental data available to date for such a structure comparison. The geometry analysis revealed that a substantial difference in shape between the conformation observed in the single crystal and that found in the protein binding site is a consistently occurring phenomenon for flexible compounds. The quantitative accuracy of these measured differences as well as of the conformational energies derived for the structures depends on the uncertainties in the original experimental data.

The typical uncertainty in atomic coordinates in small molecules determined by X-ray diffraction is on the order of 0.01 Å and this applies to most entries in the CSD.[14,32] This uncertainty is generally much larger in macromolecules,[33] and the precision of the atomic coordinates that was reported in the PDB files used for this study was typically in the range of 0.2–0.3 Å. This is, however, an *absolute* coordinate uncertainty which is generally larger than the error in the *relative* positions of two atoms connected by a covalent bond since reasonable bond length restraints[33] are usually applied during structure refinement. It follows that for RMS_{CSD} or $RMS_{Glob.Min.}$, while values below 0.3 Å may potentially be ascribed to uncertainties in the atomic coordinates used in the calculation, RMS values above 0.3 Å indicate real conformational differences between the two structures being compared. In 22 of the 27 non-rigid compounds in Table 3, RMS_{CSD} exceeds 0.3 Å, often by a considerable amount and it seems clear that the conformation in the CSD differs significantly from that in the protein-bound species, the PDB structure. No single compound with five or more rotatable bonds showed a match with RMS < 0.5 Å between the CSD and the PDB conformations. These observations suggest that much care should be taken in the use of single crystal X-ray structures of flexible molecules in connection with drug design.

The other result from the geometry comparison is that the protein-bound structures do not correlate well with the global minimum energy conformations of flexible molecules and this should be another source of concern for those engaged in molecular modeling. The problem in this case is that global minimum energy structures are usually developed with an isolated molecule in vacuum. Under such conditions, molecules are unable to form external hydrogen bonds, to solvent, protein or other molecules and as a result, they often form intramolecular hydrogen bonds. This causes the molecules to collapse into folded conformations which bear little resemblance to anything encountered in an environment of interest such as a solution or a protein active site. In Figure 6, the single crystal form of methotrexate (**24**) can be seen to be quite different from the protein-bound form (RMS = 2.60 Å) which, in turn, is different from the global minimum energy structure (RMS = 3.47 Å). This suggests that the global minimum energy structure of a flexible molecule determined in vacuum is probably even less relevant for those seeking structures which can bind in enzyme sites. We have shown previously[34] that re-determination of the minimum energy vacuum structure in a water environment often leads to a quite different structure and the results adduced here confirm this observation. Figure 6 also illustrates the additional finding that, for flexible compounds, the protein-bound conformation is often not a local minimum on the molecule's energy surface calculated in vacuum.

The influence of the uncertainties in macromolecule X-ray structure determination on the conformational *energies* is somewhat more difficult to assess, since this requires knowledge of the uncertainty in the experimental torsion angles that were used to derive the QUANTA models. Neither the resolutions of the PDB structures, listed in Table 2, nor the average absolute coordinate uncertainties mentioned above do easily translate into torsional uncertainties, and surprisingly little work seems to have been done on this question. In an analysis[35] of several hundred PDB structures, a clear dependency of the standard deviations of the protein Φ, Ψ, X_1 and X_2 torsions on the reported resolutions was found. However, even at the best resolutions of 1.3 Å or less, a standard deviation on the order of 10°–15° remained, and extrapolation of the best fit lines to a hypothetical resolution of 0.0 Å yields a residual standard deviation between 2.9° and 9.3° for those protein torsions. These residuals are likely to represent the conformational variability present in the macromolecule structure for each of those torsions, and have to be subtracted from the measured standard deviations to yield the true experimental torsional errors. Calculating these for our data set's average resolution of 2.25 Å and averaging them over all torsion angles listed above yielded a value of about 14°. Since no information on experimental errors specific to the *ligands* in protein–ligand complexes seems to be available, this value, which is based on a worst-case scenario of 0.0 Å resolution, was taken as the average torsional uncertainty in the PDB structures of the small molecules analyzed in this study.

This value of 14° was compared to the torsion angle changes observed during minimization of 21 different PDB conformers which had shown energy changes (local energies) of 4 kcal mol^{-1} or higher (average local energy 12.6 ± 6.6 kcal mol^{-1}). A total of 142 torsion angles involving only heavy atoms were measured before and after minimization. Out of 142 torsion angle changes, 98 fell outside the first bin of torsional deviations ranging from 0° to 14°, which means that 69% of the changes are greater than the average experimental torsional uncertainty as derived above. A

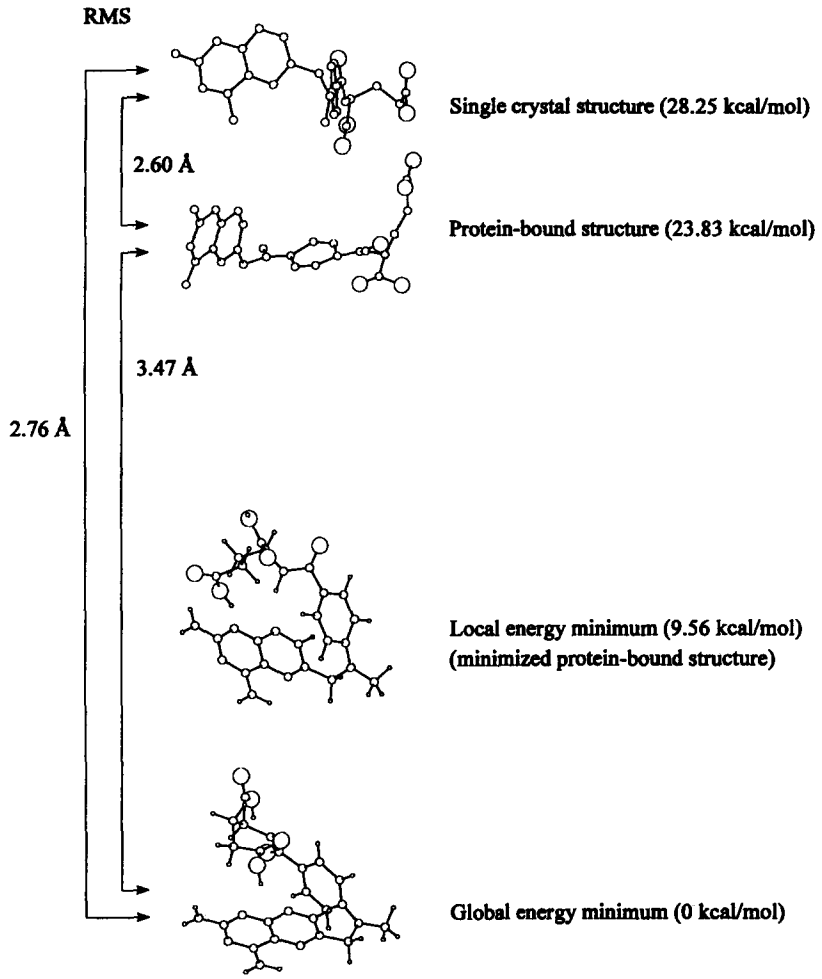


Figure 6. Conformational changes of methotrexate (**24**). From top to bottom: single crystal structure (CSD ref. code DOJZAD01); methotrexate bound to *E. coli* dihydrofolate reductase (first occurrence in PDB entry 4dfr, segment id 'AHET'); CHARMm-minimized model of the above protein-bound structure; CHARMm global energy minimum conformation. Vertical placement approximately proportional to energy. The energy for the local minimum structure (9.56 kcal mol^{-1}) corresponds to ΔE_{global}–ΔE_{local} in Table 4. Also shown are RMS differences between conformations. Large circles are oxygen atoms, medium-size circles are carbon or nitrogen atoms, small circles are hydrogen atoms (not shown for the CSD and PDB structures, no hydrogens present for methotrexate in either database).

fraction of 44% of the changes fell outside the first two bins, i.e. was 30° or larger. The average torsional change was 34.4° ± 28.9°. For all 21 conformers analyzed, the variation in the individual torsion that had changed the most was 20° or more, with values observed of up to 118°, and an average of 72° ± 29°. These findings strongly suggest that the local and, even more so, the global energies were generally well above the experimental uncertainties.

A case where lower resolution seems to be linked with higher—probably unrealistically high—conformational energies is citrate (**11**), where the lowest resolution (2.8 Å) structures, obtained from PDB entry 8ldh, show the highest energies. Low resolution did however not inevitably entail high CHARMm energies, since three other structures with exactly the same resolution, two PDB structures of malonic acid (**22**), and pyridoxamine-5'-phosphate (**29**) showed either very low energies or intermediate values, respectively.

Some of the sugars [e.g. β-D-glucose (**17**)] showed energy differences between the CSD conformations with minimized and unminimized hydroxyl torsions of up to 22 kcal mol^{-1}. Since practically all torsional conformational freedom in sugars resides in hydroxyl torsions, it is possible that these high energy differences reflect true conformational changes occurring during crystallization—and such changes may also accompany the lodging of sugars in a protein binding site, which is suggested by the results obtained for the protein–sugar complex of β-D-glucose (**17**) (PDB entry 3gbp, Table 4).

The finding that flexible molecules can be found in a protein binding site in a conformation well above the global vacuum energy minimum were supported by the semi-empirical calculations performed.

Correlating the conformational changes with the flexibility of the ligands revealed a distinct jump in dissimilarity between the protein-bound and both the single-crystal (Fig. 1) and the global minimum conformation (Fig. 2) for molecules whose rotor count exceeded four. Likewise, very few global conformational energies of less than 10 kcal mol^{-1} were found for such 'truly flexible' compounds (Fig. 3), and a distinct jump in the local conformational energies, to values of 5 kcal mol^{-1} or higher in most cases, was also observed (Fig. 5) for compounds with five or more rotors. It appears that the bound ligand conformations, which are determined by the highly structured, non-isotropic environment of the protein binding site, are virtually random points[8,36–38] (within the energy limits found) when located on the energy surface calculated for the isotropic vacuum environment. Similar observations have been made for single crystal structures.[10]

The hydrogen bond is widely regarded[25] as being the single most important intra- and intermolecular cohesive force and a major contributor of non-covalent interaction energy in biological systems. Comparison of the conformational energies with the number of ligand H-bond centers as well as with the H-bonds formed in the complexes suggests that ligands can 'use up' substantial parts of the energy gained from the formation of hydrogen bonds[28] for the conformational deformation necessary for accommodation in the binding site. An example of an array of 16 H-bonds anchoring a ligand in the binding site in a specific conformation is given in Figure 7 depicting methotrexate bound to *L. casei* dihydrofolate reductase.

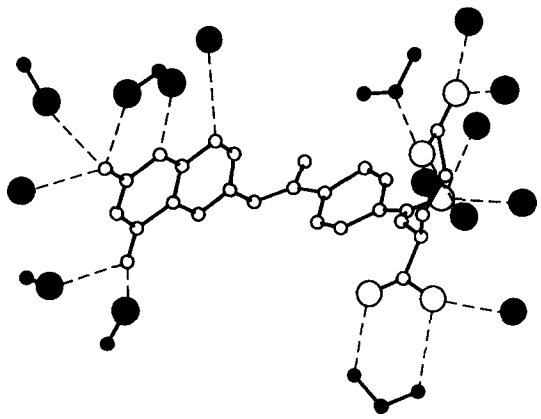


Figure 7. Methotrexate (**24**) in its binding site. Depicted is methotrexate complexed with *L. casei* dihydrofolate reductase (PDB entry 3dfr). Open circles connected by thick lines: methotrexate molecule. Filled circles: binding site atoms. Large circles are oxygen atoms, small circles are carbon or nitrogen atoms. Hydrogens are not shown (not present in the PDB file). Dashed lines: possible hydrogen bonds between methotrexate and binding site atoms. Isolated binding site oxygens are water molecules. Binding site oxygens connected to a carbon (small filled circle) are carbonyl or hydroxyl oxygens. One carboxyl group is also shown. All binding site hydrogen bond centers at a distance of 3.4 Å or less from the nearest methotrexate atom were included in the drawing. The methotrexate molecule has a bent shape in 3dfr. The reader faces the concave side of the molecule, the pteridine and the L-glutamate moieties pointing out of the paper plane.

In the context of drug development, the findings of this study suggest that it would be a reasonable approach, for example, to fit a flexible molecule to a pharmacophore without performing any energy minimization first, and only then to test, in a second step, whether a possible hit falls in the allowed energy range. Procedures along these lines are being incorporated, e.g. in a growing number of 3D database building and searching systems.[39–42]

Conclusions

Flexible molecules are deformed when binding to proteins. This deformation is a general phenomenon and the reason for it is presumed to lie to a good extent in the molecule's search for hydrogen bonds to the protein to replace the solute–solvent hydrogen bonds which are lost as the molecule enters a binding site. The degree of deformation appears to depend somewhat upon the number of rotors in the ligand. For molecules with five or more rotatable bonds, the crystal structure does not

generally well represent the protein-bound conformation in shape. Structures of flexible compounds determined by single crystal X-ray crystallography are thus less useful in the context of drug design than has generally been assumed.

The global minimum energy conformation calculated in vacuum is usually an even worse substitution for the bound structure of a flexible molecule. However, it is a structure that is useful to compute because it can be used to anchor an energy scale and allow recognition of high energy conformations.

As to conformational energies, the bound conformation of a flexible molecule is not usually any of the low energy minima found in vacuum. Instead, one has to be prepared to find the protein-bound conformation virtually anywhere on the energy surface within a certain range above the global minimum. The number of ligand atoms capable of forming hydrogen bonds may serve as a parameter for a first, semi-quantitative estimate of the energy that is available for deformation of the ligand.

The only solution to the problem that ligand flexibility causes in drug design lies in the use of wide conformational space when searching for bioactive conformations. Rather than use a single conformer when searching for compounds fitting a pharmacophore or when examining the binding of a ligand to an enzyme, it is important that all possible conformers be considered, at least at the outset. Elimination of high-energy conformations is entirely permissible, but restricting oneself to the conformation measured in the isolated crystal or, for that matter, the global energy minimum calculated in vacuum is clearly a dangerous practice.

Experimental

Data retrieval

A search was conducted in the 1492 PDB files (1055 full release and 437 pre-release entries in PDB v. 63) for the string 'HET ' at the beginning of a line, which denotes lines in the header section of the PDB files that list 'hetero-residues' occurring in the structure. Removal from the retrieved records of those in which the 'foreign' species was a single atom or ion (such as a halogen or a metal counter ion), a very small molecule or ion (such as sulfate or phosphate), a solvent molecule (such as methanol), or a covalently bound fragment (such as a methyl group), left a total of 197 compound names as the set of potentially useful ligands. Searching the headers for terms such as 'COMPLEX', 'LIGAND', 'BOUND', etc., yielded fewer retrievals.

The 197 compound names retrieved from the PDB were used in a text string search of the compound name fields of the CSD (using CSD's search program QUEST, v. 4.50). Hits in which the search string was only a substring of a longer chemical name were discarded, and out of a total of 90296 records in the CSD, this led to 137 retrievals, several of which represented crystal structures of the same small molecule. There were no atomic coordinates in 27 of the 137 CSD entries, which left 110 structures. These structures, which were extracted from the database in the CSD's "FDAT" format were then converted by means of a FORTRAN program[43] to a fractional coordinates format which was used as input to the molecular mechanics programs QUANTA.

Where multiple entries were found in the CSD for a single structure, an effort was made to find the 'best' entry, i.e. structure of a pure compound, free of molecules of solvation or counterions. Duplicate entries were included but all "inferior" entries were excluded. A number of the PDB entries were also excluded for various reasons, such as covalent binding between the macromolecule and the 'ligand' (such as acetylcholine in the PDB entry 'lace'), and several cases in which the structures from the PDB and CSD were chemically distinct in spite of the names applied by the databases. For example, in a number PDB entries, sugars are listed under their individual monomer names while the bound structure is actually a polysaccharide. Finally, five compounds which described ligands bound to DNA were also dropped.

The presence of unique compound identifiers such as CAS Registry Numbers (RN) in both databases would facilitate such combined searches, but only about 10% of the entries in the CSD carry CAS RN, and entries in the PDB do not usually list CAS RN.

Modeling of specific compounds

Dimethylformamide (12). All three DMF molecules present in the PDB entry 6est[44] and all five DMF molecules in 7est[44] were found in virtually the same conformation with the only torsion angle in the molecule having a value very close to 90°, a torsional energy *maximum* according to both CHARMm ($E \approx 10$ kcal mol^{-1}; Table 4) and AM1 (E = 14.39 kcal mol^{-1}; Table 5). This is hardly reconcilable with the maximum energy that would be available from optimal H-bond formation, and it seems difficult to conceive that *all* eight DMF molecules in the PDB entries would be uniformly found in their energy maximum (the CSD conformer is in fact planar). Because of this uncertainty about the validity of the experimental coordinates, only one data point, the average of all eight energies (9.75 kcal mol^{-1}), was used for DMF in Figures 3–5.

FK506 (14). For FK506, the derived structure was constructed by applying CHARMm 'NOE' (distance) constraints, with a no-force distance interval of zero ($R_{min} = R_{max} = 0.0$ Å) and a force constant in either direction (K_{min}, K_{max}) of 60 kcal mol^{-1} Å^{-2} (QUANTA parameters: 'Deviation high/low' = 0.1 Å, scale factor = 1.0, temperature = 300 K). A bond stretching energy of

2.2 kcal mol^{-1} was produced by this method. The RMS deviation between the heavy atom coordinates of the derived and the original PDB structure was 0.068 Å.

p-*Hydroxybenzoic acid (18).* Although possessing one rotatable bond according to our definition, *p*-hydroxybenzoic acid is essentially a rigid, planar compound because of resonance between the carboxyl group and the aromatic ring. This is borne out by the conformational energies of practically zero for all three structures analyzed (Table 4), which were therefore excluded from the analyses and plots (Figs 3–5) of conformational energies of *flexible* compounds.

Sucrose (32). For the protein-bound molecule of sucrose from the PDB entry 1r1a[45], achieving an even moderately good fit between the derived model of the bound structure and the coordinates in 1r1a proved to be completely irreconcilable with the energies available from the four H-bonds counted in the protein–ligand complex. The conformational energy of 57 kcal mol^{-1} (Table 4) was obtained for a derived structure that showed a RMS deviation of 0.96 Å from the experimental coordinates. Manually adjusting the bond angles in the model improved the fit to an RMS value of 0.39 Å but caused the energy to increase to 99 kcal mol^{-1}. The sucrose molecule in the protein X-ray structure had been posited for an unexplained electron density found in that region, and had been termed 'putative sucrose' in the original publication.[45] The PDB entry 1r1a is also the structure with the lowest resolution (3.2 Å) and the highest *R*-factor (0.293) in the compound set. From the data obtained in this study, it seems questionable that the sucrose molecule is present in the complex in the conformation/orientation given in the PDB entry. Because of this probable unreliability of the conformational energies, no data point for sucrose was used in Figures 3–5 and all analyses involving conformational energies.

Acknowledgments

We thank Dr Kevin P. Cross for providing us with the GS and GSN flexibility indices of the compounds used in this study. We are grateful to Dr Victor E. Marquez for careful reading of the manuscript and many helpful suggestions. We thank Dr Konrad F. Koehler for valuable discussions.

References

1. Fischer, E.; Thierfelder, H. *Chem. Ber.* **1894**, *27*, 2031.

2. Fischer, E. *Chem. Ber.* **1894**, *27*, 2985.

3. Koshland, Jr D. E. *Proc. Natl Acad. Sci. U.S.A.* **1958**, *44*, 98.

4. Jorgensen, W. L. *Science* **1991**, *254*, 954.

5. Ricketts, E. M.; Bradshaw, J.; Hann, M.; Hayes, F.; Tanna, N.; Ricketts, D. M. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 905.

6. Rice, K. G.; Wu, P.; Brand, L.; Lee, Y. C. *Curr. Opin. Struct. Biol.* **1993**, *3*, 669.

7. Moodie, S. L.; Thornton, J. M. *Nucleic Acids Res.* **1993**, *21*, 1369.

8. Klebe, G. In: *Structure Correlation*, Vol. 2, pp. 543–603, Bürgi, H. B. Dunitz, J. D. Eds; Verlag Chemie; Weinheim, 1994.

9. Klebe, G. *J. Mol. Biol.* **1994**, *237*, 212.

10. Pranata, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1991**, *113*, 9483.

11. Andrews, P. R.; Craik, D. J.; Martin, J. L. *J. Med. Chem.* **1984**, *27*, 1648.

12. Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Crystallogr.* **1979**, *B35*, 2331.

13. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, Jr. E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535.

14. Stout, G. H.; Jensen, L. H. *X-Ray Structure Determination. A Practical Guide*; Wiley; New York, 1989.

15. Fisanick, W.; Cross, K. P.; Rusinko, III A. *Tetrahedron Comput. Methodol.* **1990**, *3*, 635.

16. QUANTA is provided by Molecular Simulations, Inc., 16 New England Executive Park, Burlington, MA 01803-5297, U.S.A.

17. Silicon Graphics, Inc., 2011 N. Shoreline Blvd., Mountain View, CA 94039-7311, U.S.A.

18. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.

19. *QUANTA Version 3.3 Parameter Handbook*, pp. 1–172, Molecular Simulations, Inc. Burlington, MA 01803, 1992.

20. Derreumaux, P.; Zhang, G.; Shlick, T.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 532.

21. Veal, J. M.; Wilson, W. D. *J. Biomol. Struct. Dyn.* **1991**, *8*, 1119.

22. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

23. Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209.

24. Stewart, J. J. P. *J. Comput. -Aided. Mol. Design* **1990**, *4*, 1.

25. Jeffrey, G. A.; Saenger, W. *Hydrogen Bonding in Biological Structures*, pp. 1–569, Springer; Berlin, 1991.

26. Clark, D. E.; Willett, P.; Kenny, P. W. *J. Mol. Graphics* **1993**, *11*, 146.

27. Bohacek, R. S.; McMartin, C. *J. Med. Chem.* **1992**, *35*, 1671.

28. Andrews, P. R.; Tintelnot, M. In: *Comprehensive Medicinal Chemistry*, pp. 321–347, Hansch, C. Ed.; Pergamon Press; Oxford, 1990.

29. Chakrabarti, P. *J. Mol. Biol.* **1993**, *234*, 463.

30. Vermersch, P. S.; Tesmer, J. J.; Quiocho, F. A. *J. Mol. Biol.* **1992**, *226*, 923.

31. Williams, D. H.; Searle, M. S.; Mackay, J. P.; Gerhard, U.; Maplestone, R. A. *Proc. Natl Acad. Sci. U.S.A.* **1993**, *90*, 1172.

32. Allen, F. H. In: *Accurate Molecular Structures: Their Determination and Importance*, pp. 355–378, Domenicano, A.; Hargittai, I. Eds; Oxford University Press; Oxford, 1992.

33. Laskowski, R. A.; Moss, D. S.; Thornton, J. M. *J. Mol. Biol.* **1993**, *231*, 1049.

34. Milne, G. W. A.; Nicklaus, M. C.; Hodoscek, M. *J. Mol. Struct.* **1993**, *291*, 89.

35. Morris, A. L.; MacArthur, M. W.; Hutchinson, E. G.; Thornton, J. M. *Proteins* **1992**, *12*, 345.

36. Klebe, G.; Mietzner, T. *J. Comput.-Aided Molec. Design* **1994**, *8*, 583.

37. Chang, Y.-T.; Loew, G. H.; Rettie, A. E.; Baillie, T. A.; Sheffels, P. A.; Ortiz de Montellano, P. R. *Int. J. Quant. Chem.: Quant. Biol. Symp.* **1993**, *20*, 161.

38. Loew, G. H. *Second Annual Meeting on Development of Small Molecule Mimetic Drugs*, Philadelphia, PA, April 11–12, 1994.

39. Murrall, N. W.; Davies, E. K. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312.

40. Leach, A. R.; Prout, K. *J. Comput. Chem.* **1990**, *11*, 1193.

41. Moock, T. E.; Henry, R. R.; Ozkabak, A. G.; Alamgir, M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 184.

42. Hurst, T. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190.

43. Hendrickson, M. A.; Nicklaus, M. C.; Milne, G. W. A.; Zaharevitz, D. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 155.

44. de la Sierra, I. L.; Papamichael, E.; Sakarellos, C.; Dimicoli, J. L.; Prangé, T. *J. Mol. Recognit.* **1990**, *3*, 36.

45. Kim, S. S.; Smith, T. J.; Chapman, M. S.; Rossmann, M. G.; Pevear, D. C.; Dutko, F. J.; Felock, P. J.; Diana, G. D.; McKinlay, M. A. *J. Mol. Biol.* **1989**, *210*, 91.

(*Received in U.S.A.* 14 *November* 1994; *accepted* 18 *January* 1995)